Global Local Folding of the Human Transcriptome

Undergraduate Research Thesis

Presented in Partial Fulfillment of the Requirements for Graduation with "Honors Research Distinction in Microbiology" in the Undergraduate Colleges of The Ohio State University

by
James Li

The Ohio State University
May 2018

Project Advisor: Professor Peter White[1]
Defense Committee Advisors: Professor Birgit Alber[2] & Professor Ralf Bundschuh[3,4,5,6]
Major Advisor: Professor Natividad Ruiz[2]

Department of Pediatrics[1], Department of Microbiology[2], Department of Physics[3], Department of Chemistry & Biochemistry[4], Department of Hematology[5], and Department of Internal Medicine[6]

Abstract

Analyzing sequence variants for disease has largely been based on the effects of missense mutations on predicted changes to protein function. Previous *in silico* RNA folding studies suggest that selection in humans and mammals may have been influenced by mRNA secondary structure in certain genes. However, the connection between RNA folding and genetic diseases has not been fully established at the level of an entire transcriptome. Therefore, we performed whole transcriptome analysis to ascertain the effects of single nucleotide polymorphisms (SNPs) on local RNA folding. We aimed to (1) build a cloud-based big data pipeline to procure RNA folding statistics for every possible polymorphism in the known human transcriptome (~0.5 billion variants), (2) utilize population allele frequencies from 138,632 patients as well as mammalian conservation scores to determine if there was constraint on SNPs causing large RNA disruptions, thereby supporting our hypothesis that RNA stability/structure may play a role in disease and (3) develop a tool and composite score to rapidly analyze patient genomes for highly disruptive SNPs. For every position in all known RefSeq mRNA transcript sequences, we generated flanking sequences (101 nucleotides each) corresponding to the reference allele and the three possible alternate alleles. Next, we used the ViennaRNA Package to obtain 10 RNA folding disruption metrics for each possible variant (445,740,246 total SNPs). We then sorted the SNPs for each of their ten RNA folding metrics and binned these SNPs based on the percentiles of each of their metric values. Metric bins with higher RNA disruption values had a larger proportion of SNPs with an allele frequency equal to zero, compared to bins with lower RNA disruption values. Similarly, median and mean GERP++ scores were greater for higher disruption bins than lower disruption bins. The correlation of increased RNA disruption values with both decreased allele frequencies and increased GERP++ scores at the level of the whole human transcriptome, suggests that RNA folding plays an important role in human health and disease.

Acknowledgments

I'd like to first and foremost thank Professor Peter White for being my super wonderful PI and chairing my defense committee. Academically, he has taught me a lot and helped tremendously to develop my critical thinking and writing skills. On top of just being a mentor intellectually, Professor White has provided me with tons of insights into life and has generously taken lots of time out of his schedule to support me in many other aspects - including giving me a bike for Pelotonia when I didn't have one! These few lines cannot fully express how much I am indebted to all his mentoring and support. I'd also like to thank my lovely defense committee members, Professors Birgit Alber and Ralf Bundschuh for graciously taking the time to attend my defense and review my thesis.

I'd like to thank Dr. Jeffrey Gaither for his super invaluable mentoring of my project. He is so supportive of me and has helped me a ton in my work by taking the time to make sure I understand concepts that I may have trouble grasping. I'd also like to thank Mr. Benjamin Kelly for mentoring and teaching me how to use cloud-computing technologies with a neat genome alignment project. I'd also like to thank Mr. Grant Lammi for his tremendous work in scaling the RNA folding project's pipeline, as well as Mr. David Gordon, Mrs. Harkness Kuck, and Mr. James Fitch for all their contributions to this project.

I'd also like to thank all my past mentors for their guidance and skills they have taught me. I'd like to thank Professor Biao Ding and Mr. Jian Wu for instilling a huge love of genetics and RNA biology in me, as well as spending a lot of time to make sure I understood the literature I was exposed to during my freshman year. I'd like to thank Dr. Andrew Johnson for his mentoring during the summer after my freshman year when he introduced me to population genetics and the potential link between RNA structure and human genetic markers. I'd like to thank Dr. John Malas who mentored me during high school and gave me my first real appreciation for the concept of statistical dimensionality, which let me see the potentially applicability of radar signature statistical analysis to RNA biology. I'd like to thank Dr. Lorraine

Wallace for mentoring and providing me with the amazing opportunity to present my work at the 2017 University of São Paulo International Symposium of Undergraduate Research in Brazil.

I'd like to thank the Morrill Scholars Program here at Ohio State, as well as the Pelotonia Fellowship Program for generously helping to finance my undergraduate career and providing me with numerous opportunities to share and present my research.

I'd like to thank all my friends for their support and my bestfriend Kevin Lin for always sticking with me and rooting for me throughout my life. And lastly, I'd like to thank my dear family for their never ending love and support of me to pursue my dreams.

Table of Contents

I. Abstract

II. Acknowledgments

III. Table of Contents

1. Introduction
    - 1.1. Types of Mutations
    - 1.2. Classes of SNPs
    - 1.3. Sequence Variant Analysis
    - 1.4. RNA Folding and Allele Frequency
    - 1.5. Hypothesis
    - 1.6. Objectives
    - 1.7. A Systematic Whole Transcriptome Analysis of RNA Folding
    - 1.8. Transforming Our Analysis into a Measure of Deleteriousness
2. Tools
    - 2.1. The ViennaRNA Package
    - 2.2. Liftover
    - 2.3. Genome Aggregation Database (gnomAD)
    - 2.4. SnpEff
    - 2.5. Genomic Evolutionary Rate Profiling (GERP)
    - 2.6. Shiny in RStudio
3. Materials and methods
    - 3.1. Retrieving Refseq Transcripts
    - 3.2. Generating Flanking Sequences
    - 3.3. Calculating Folding Statistics with the ViennaRNA Package
    - 3.4. Accounting for Reference Assemblies
    - 3.5. Joining Our Dataset with gnomAD
    - 3.6. Annotating with SnpEff
    - 3.7. Calculating p-values for Metrics

3.8. Developing Composite Scores
3.9. Analysis of Human Population Constraint and Mammalian Conservation
3.10. Building SnpRFC
4. Results
4.1. Whole Transcript Results
4.2. 5’ Untranslated Region (5’ UTR)
4.3. 3’ Untranslated Region (3’ UTR)
4.4. Synonymous
4.5. Missense
4.6. SnpRFC: “SNP mRNA Folding Consequences in Humans”
5. Discussion
6. Conclusions and Future Work
7. References
8. Supplementary Figures
9. Figure and Table Appendix

1. Introduction

The widespread adoption of next-generation sequencing (NGS) for the study of human genetic disease over the past decade has generated a plethora of sequencing data. As opposed to Sanger sequencing, which utilizes dideoxynucleotide termination and subsequent capillary electrophoresis to generate single reads of up to 900 bp, NGS technologies use massively parallelized sequencing approaches, generating billions of sequence reads and enabling rapid sequencing of entire human genomes. This increase in our ability to generate sequence data provides us the opportunity to procure high quality sequence variant data for individuals, disease cohorts, and entire populations.

1.1 Types of Mutations

Sequence variants refer to the general category of changes to the genome that may or may not alter a particular phenotype. The major classes of genetic variants include single nucleotide polymorphisms (SNPs), small insertions and deletions (indels), and larger structural variations (SVs) such as inversions, translocations, and copy number variations (CNVs). In this thesis, we will focus specifically on the most prevalent form of genetic variation, SNPs, which refer to the substitution of a single base for another base at a particular position, resulting in either synonymous or non-synonymous variants. Non-synonymous changes can either be missense (a change in the encoded amino acid) or nonsense (loss of a start codon, or introduction of a premature stop codon resulting in a truncated protein). Due to redundancy in the genetic code, synonymous changes do not result in a change in the amino acid encoded by a given codon. The majority of genetic variants are considered benign and together constitute normal genetic variation, but the handful of variants that contribute to disease development and progression are classified as being pathogenic.

1.2 Classes of SNPs

Two fundamental classes of SNPs are transversion and transition mutations. Transversion mutations refer to the exchange of a pyrimidine for a purine or a purine for a pyrimidine (i.e. cytosine to guanine or adenine to cytosine, respectively), while transition mutations refer to the exchange of a pyrimidine for a pyrimidine or a purine for a purine (i.e. cytosine to thymine or adenine to guanine, respectively).

In addition to their underlying biochemical change, SNPs can also be partitioned by their genomic location into coding and noncoding variants. For messenger RNAs (mRNAs), a coding variant refers to a SNP that falls between the start and stop codon codon of a given transcript, while a non-coding variant refers to a SNP that lies outside the transcript's open reading frame. Coding variants can be further distilled into synonymous and non-synonymous mutations, which refer to whether the SNP causes a change in the eventually coded amino acid. Noncoding variants in mRNA can be categorized as 5' and 3' UTR variants depending on which tail they lie within. Non-coding variants are also found in non-coding RNAs (such as microRNAs or long non-coding RNAs) and throughout the non-exonic regions in the genome.

1.3 Sequence Variant Analysis

Currently, many bioinformatics tools that are used to predict the effects of SNPs such as PolyPhen and SIFT look at functional impacts on the resulting proteins (Adzhubei *et al.* 2010 & Kumar *et al.* 2009). While these tools have utility in quickly procuring the disease-related effects that variants may have on proteins, they miss out on other potential etiologies. Studies in the past two decades have noted that synonymous variants (sSNPs) may also play a crucial role in gene expression and protein translation, through influencing mRNA stability/structure, mRNA splicing and maturation, as well as protein translation rates and folding (Soussi *et al.* 2017; Holmila *et al.* 2003; Raponi *et al.* 2010; Sauna *et al.* 2011; Gartner *et al.* 2013; Supek *et al.* 2014; Gotea *et al.* 2015).

Despite the recent emergence of disease-associated SNPs that do not conspicuously change protein function, several public databases such as TCGA's cBioPortal do not contain synonymous SNPs (Soussi *et al.* 2017), despite the fact that current estimates suggest sSNPs could account for 50% of driver mutations in cancer.

1.4 RNA Folding and Allele Frequency

In entertaining the effects of RNA folding on human disease, several studies have attempted to correlate RNA folding with population allele frequencies. In a study examining 34,557 SNPs in ~12,450 Refseq genes, Johnson showed that SNPs with larger predicted minimum free energy changes to local mRNA structure resulted in lower allele frequencies reported by dbSNP (Johnson *et al.* 2011). Similarly, Vilmi examined 96 SNPs in 22 tRNA genes and found that allele frequencies among 912 individuals were lower for variants predicted to be more disruptive (Vilmi *et al.* 2005). Though these early studies suggested constraint in allele frequencies may correspond to increased RNA folding disruptions, the connection between RNA folding and genetic diseases has not yet been fully established at the level of the whole human transcriptome.

1.5 Hypothesis

If RNA folding plays a role in human health and disease, then SNPs that cause large RNA folding disruptions should have constrained population allele frequencies and conservation scores.

1.6 Objectives

The purpose of this thesis is three-fold. We aimed to (1) build a cloud-based big data pipeline to procure RNA folding statistics for every possible polymorphism in the known human transcriptome (~0.5 billion variants), (2) utilize population allele

frequencies from 138,632 patients as well as mammalian conservation scores to determine if there was constraint in SNPs resulting in large RNA disruptions, and (3) introduce a tool and composite score to rapidly annotate SNPs with RNA folding statistics, population allele frequencies, and conservation scores.

1.7 A Systematic Whole Transcriptome Analysis of RNA Folding

There has yet to be a systematic whole transcriptome analysis of RNA folding. Our approach was to calculate local RNA folding statistics for every possible polymorphism in the human mRNA transcriptome and then tie in population allele frequencies from the Genome Aggregation Database (gnomAD) for each polymorphism to ultimately examine the relationship of both population allele frequencies and mammalian conservation scores with RNA folding disruption metrics.

1.8 Transforming Our Analysis into a Measure of Deleteriousness

As previously mentioned, tools such as SIFT and PolyPhen analyze non-synonymous mutations to predict their deleteriousness to protein function and production. We thus developed a tool - SnpRFC or "SNP mRNA Folding Consequences in Humans" - to assign every SNP a set of RNA folding disruption scores. We designed SnpRFC to be a comprehensive tool that calculates 10 different RNA folding statistics and ties in population allele frequencies and conservation score annotations for a given set of SNPs. Moreover, p-values for RNA folding disruption calculated by SnpRFC are based on the disruption of a particular SNP compared to the disruptions of all possible SNPs in the human mRNA transcriptome - this is our attempt to reach a more biologically intuitive p-value, as opposed to comparisons to disruptions obtained from a stochastically generated null population. Ultimately, SnpRFC reports composite disruption scores for a given SNP's influence on RNA folding.

2. Tools

The following SNP annotation tools were used to generate RNA folding metrics, as well as retrieve genomic coordinates for reference genomes, population allele frequencies, transcript locations, mutation types, and conservation scores.

2.1. The ViennaRNA Package

The ViennaRNA package is a suite of programs that computationally predicts RNA secondary structures of sequences input by a user (Lorenz *et al.* 2008). Several aspects of RNA folding predicted by Vienna include free energy, specific heat, ensemble diversity, base pairing probabilities, structures of sequences, as well as distance and structural conservation between sequences. The ViennaRNA Websuite is an accessible, online platform of RNA folding algorithms that compute metrics such as free energy, positional entropy, and base pairing probabilities, in addition to generating mountain plots and visual representations of folded RNA sequences (Gruber *et al.* 2008).

2.2. Liftover

The Batch Coordinate Conversion or *liftOver* tool developed at the University of California, Santa Cruz, is used to interconvert genomic coordinates and annotations between reference genomes (Kent *et al.* 2003).

2.3. Genome Aggregation Database (gnomAD)

The Genome Aggregation Database or gnomAD is a global project that has compiled whole genome and exome sequence data from a total of 138,632 unrelated human individuals (Lek *et al.* 2016). gnomAD is an update to the Exome Aggregation Consortium, which only included exome sequence data. Variants are annotated with their

respective population allele frequencies and sequence coverage. A total of 123,136 exomes and 15,496 genomes are included in gnomAD.

2.4. SnpEff

SnpEff is a software tool that utilizes genomic positions of variants to ascertain their effects on a given transcript: examples include silent, non-synonymous, frameshift, and start/stop codon gain/loss mutations, as well as subregions the SNPs lie in on transcripts such as 5'/3' UTRs and the coding region (Cingolani *et al.* 2012).

2.5. Genomic Evolutionary Rate Profiling (GERP)

GERP operates on the premise that purifying selection can be tracked by the lack of substitutions in a region of a given genome (Cooper *et al.* 2005). GERP++ scores at each positions rely upon what the authors term as "rejected substitutions," or "the number of substitutions expected under neutrality minus the number of substitutions 'observed' at the positions" (Davydov *et al.* 2010). We used GERP++ scores as our study's primary mammalian conservation score.

2.6. Shiny in RStudio

Shiny is a package in RStudio that is used to create apps. These apps can be web-based, standalone, or interfaced with other platforms.

3. Materials and methods

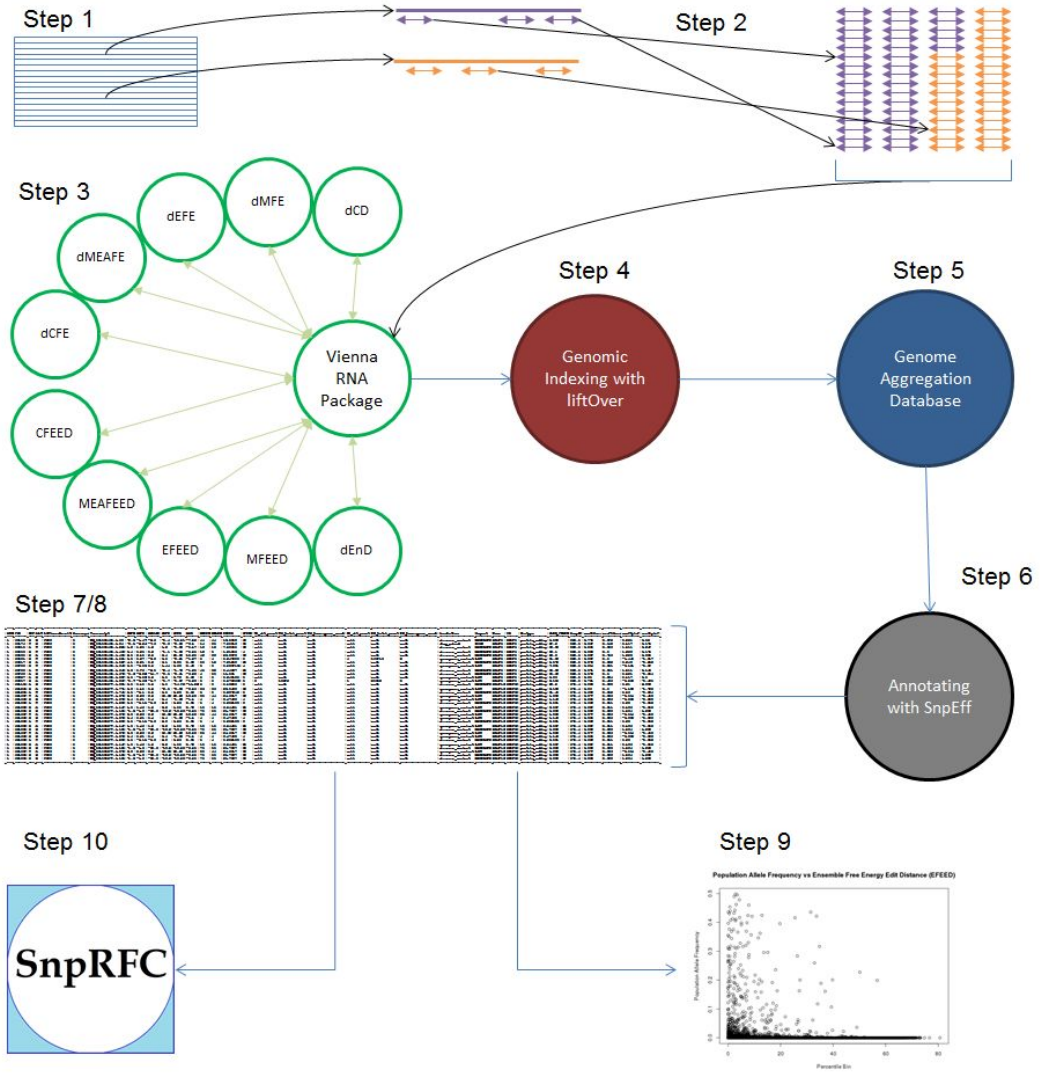


Figure 1: Diagram of the Overall Pipeline

Step 1: Retrieving Refseq Transcripts
Step 2: Generating Flanking Sequences
Step 3: Calculating Folding Statistics with Vienna
Step 4: Accounting for Reference Assemblies
Step 5: Joining Our Dataset with gnomAD
Step 6: Annotating with SnpEff
Step 7: Calculating p-values for Metrics
Step 8: Developing Composite Scores
Step 9: Analysis of Human Population Constraint and Mammalian Conservation
Step 10: Building SnpRFC

3.1. Retrieving Refseq Transcripts

NCBI Refseq Release 81 transcript sequences were retrieved on March 25, 2017 from an online repository (ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/). Transcript sequences corresponded with human reference genome build GRCh38.

3.2. Generating Flanking Sequences

For every position of each transcript, we generated four 101 nucleotide flanking sequences corresponding to one reference and three possible alternate alleles at each site. 101 nucleotide windows were used in correspondence to prior studies suggesting that a 101-151 base frame was most ideal for MFE prediction analysis (Hamasaki-Katagiri *et al.* 2017). For SNPs within 50 bases of the start or end of a transcript, the respective first or last 101 bases were retrieved to generate flanking sequences.

3.3. Calculating Folding Statistics with the ViennaRNA Package

We utilized the following three programs from the ViennaRNA package to calculate our RNA folding statistics: RNAfold, RNAdistance, and RNApdist. RNAfold -p -MEA --noPS < wildType.fasta > output_wT.fasta and RNAfold -p -MEA --noPS < SNP.fasta > output_SNP.fasta computed the:

- change in minimum free energy (dMFE)
- ensemble free energy (dEFE)
- free energy of the centroid structure (dCFE)
- free energy of the maximum expected accuracy structure (dMEAFE)
- structural ensemble diversity (dEnD)
- distance of the centroid to the ensemble of structures (dCD)
- minimum free energy secondary structure

- ensemble-weighted secondary structure
- maximum expected accuracy secondary structure, and
- centroid secondary structure of the wild type and SNP sequences.

The six metrics above had their absolute values taken to obtain magnitudes of free energy and diversity disruptions. RNAdistance < wildTypeandSNP_secondarystructures.txt > output_wTandSNP.distance calculated the:

- minimum free energy secondary structure edit distance (MFEED)
- maximum expected accuracy structure edit distance (MEAED), and
- centroid secondary structure edit distance (CFEED).

RNApdist < wildTypeandSNP_sequences.fasta > output_wTandSNP.pdist computed the:

- distance between thermodynamic ensembles of wild type and SNP sequences (EFEED).

3.4. Accounting for Reference Assemblies

The transcript database we utilized was mapped to the GRCh38 reference, yet gnomAD was in hg19 coordinates. As such, Picard liftOver was used to convert the GRCh38 coordinates of our transcripts into hg19 coordinates in order to retrieve population allele frequencies from gnomAD (http://broadinstitute.github.io/picard/index.html).

3.5. Joining Our Dataset with gnomAD

Each SNP was annotated with whole exome (EX) and whole genome (WG) alternate allele counts and total allele counts for each SNP/position via gnomAD. We divided alternate allele counts by total allele counts to procure population allele frequencies using the formula:

$$frequency = \frac{gnomAD\ EX\ alt\ allele\ count + gnomAD\ WG\ alt\ allele\ count}{gnomAD\ EX\ total\ allele\ count + gnomAD\ WG\ total\ allele\ count}$$

All SNPs that were not present in gnomAD, but had sufficient sequence coverage were assigned a population allele frequency equal to 0.

3.6. Annotating with SnpEff

SnpEff was implemented to retrieve genomic variant annotations and functional effect predictions. SNPs were also annotated with GERP++ scores, thereby enabling us to estimate the conservation of each position's reference allele among mammals.

3.7. Calculating p-values for Metrics

P-values were calculated for metrics of each SNP by finding the proportion of SNPs among the total population of SNPs with larger magnitudes for a given metric. For example, if a SNP has a minimum free energy value with a greater magnitude than 85% of the other SNPs (with a magnitude less than 15% of the SNPs), it would be given a p-value of 0.15. Columns containing p-values were added with the following R formula provided that rna$absFoldingDisruptionMetric represents a column in the data table with absolute values of each of the elements in one of the 10 RNA folding metrics:

$$rna[,\ pvalueofabsFoldingDisruptionMetric\ := 1 - ecdf(rna\$absFoldingDisruptionMetric)(rna\$absFoldingDisruptionMetric)]$$

3.8 Developing Composite Scores

We developed two composite scores to summarize our 10 RNA folding disruption metrics: (1) an extreme score and (2) a semi-weighted score. Extreme scores were procured by taking the p-value of the most significant metric and annotating the SNP with that p-value. Semi-weighted scores were procured by taking the p-value of the most significant metric for each of the three RNA folding properties: 1) stability (dMFE, dEFE, dMEAFE, dCFE), 2) structure (MFEED, EFEED, MEAED, CFEED), and 3)

ensemble diversity (dEnD, dCD), and then averaging those three p-values. These composite scores were incorporated into SnpRFC's output.

3.9 Analysis of Human Population Constraint and Mammalian Conservation

For each of the 10 RNA folding metrics we sorted the SNPs and then divided them into ten equally sized bins. % non-zero population allele frequency and mean or median GERP++ scores were plotted vs. these decile bins. Regression lines were fitted to each plot of population allele frequency or GERP++ score vs. RNA disruption metric decile bin and two-tailed hypothesis tests were conducted to see if regression lines slopes were significantly different from 0 (p-value threshold < 0.05). Example binned decile plots correlating EFEED disruptions to GERP++ score and population allele frequency are shown in Figure 2 below.

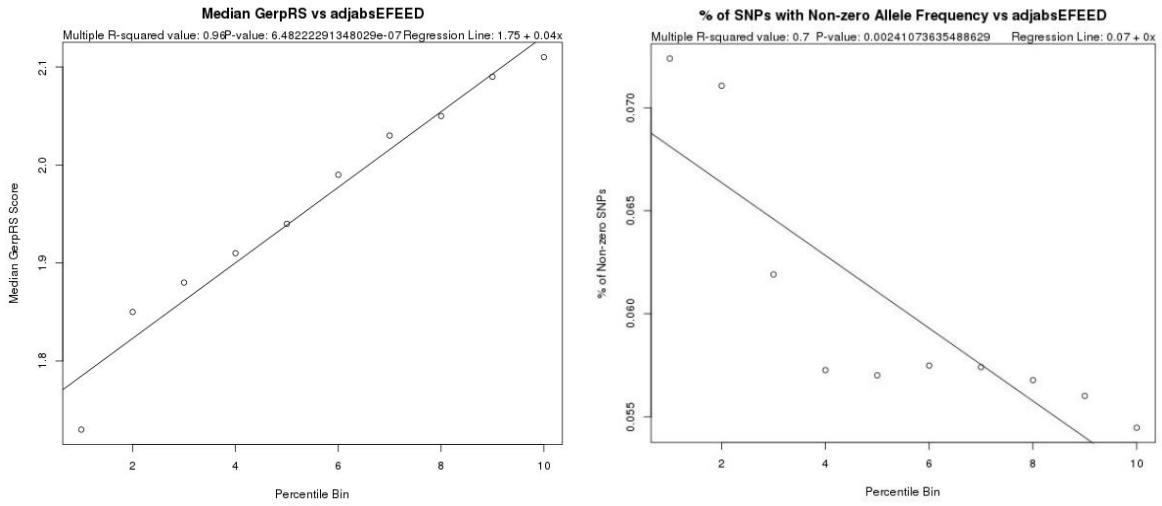


Figure 2: Example binned decile plots correlating EFEED disruptions to median GERP++ score (left) and population allele frequency (right)

If these significant regression line slopes had proper directionality (increasing disruption leading to both increasing mean and median GERP++ score, and increasing disruption leading to decreasing population allele frequency) the corresponding population allele frequency or GERP++ score vs RNA disruption metric decile bin plot

was given a score of 1 significant metric out of 1 analyzed disruption metric (for correlation to either the respective population allele frequencies or GERP++ scores). For GERP++ score vs. metric correlations with only a significant mean or only a significant median GERP++ score correlation, the plot was given a score of 0.5 significant metric out of 1 analyzed disruption metric for correlation to GERP++ score. For each RNA folding property (structure, stability, and diversity), the percentage of their corresponding metrics with both proper directionality and significant association to GERP++ score or population allele frequency was recorded by adding up their scores and then dividing by the total analyzed disruption metrics for correlation to either the respective GERP++ scores or population allele frequencies.

3.10 Building SnpRFC

SnpRFC or "SNP mRNA Folding Consequences in Humans" was developed via the Shiny package in RStudio. This app was built to search for specific SNPs (with a user input format of chromosome:positionREF>ALT [i.e. 1:69580C>A]) and retrieve their respective RNA folding statistics, population allele frequencies, GERP++ scores, and both extreme and semi-weighted composite disruption scores.

4. Results

Supplementary Figures 1-100 contain plots that analyze correlations between RNA folding metrics and both population allele frequencies and mammalian conservation scores. Summaries of the percentage of significantly constrained metrics for each of the three RNA folding properties (structure, stability, and diversity) are presented below; these summaries encompass data for all the SNPs (denoted as "Whole Transcript" results), as well as SNPs specific to certain transcript subregions. In the following figures, "GERP++" refers to the mammalian conservation metric and "AF" refers to population allele frequency in humans.

4.1. Whole Transcript Results

Figure 3/Table 1 show that an analysis of all SNPs in the transcriptome reveals that correlations between GERP++ scores and RNA folding disruptions were significant for all disruption metrics in all three RNA folding properties. Correlations between population allele frequencies and RNA folding disruptions were significant for some diversity (50%) and structure (75%) disruption metrics.

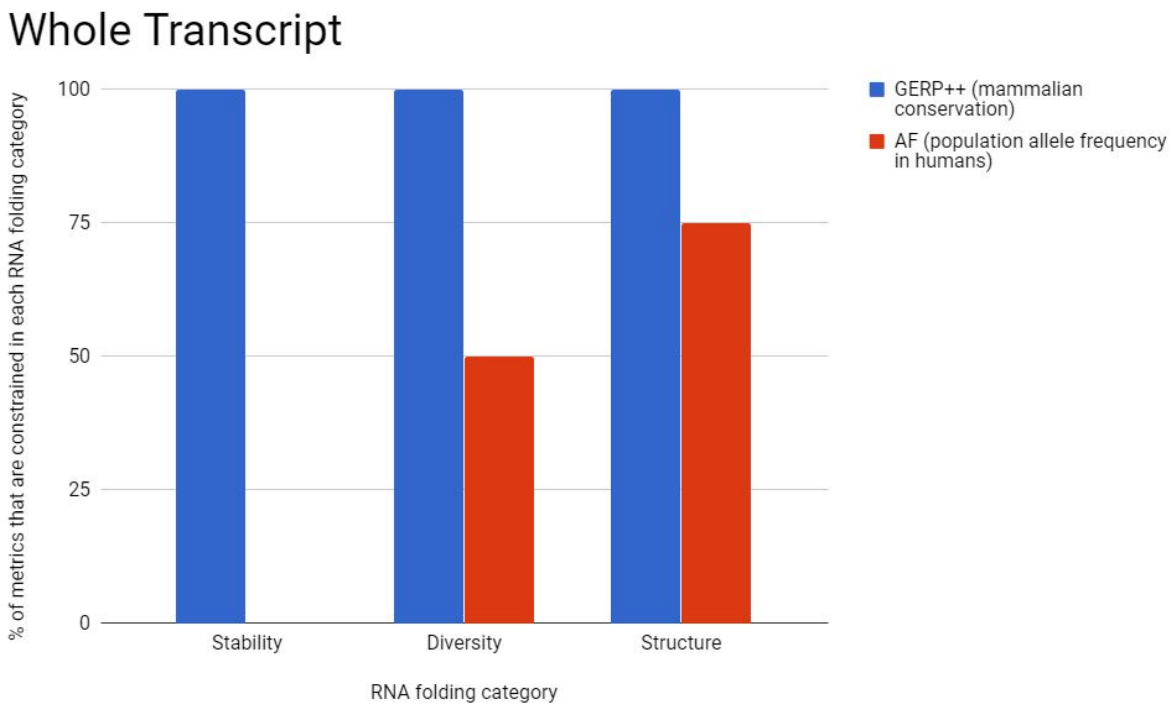


Figure 3: Percentage of disruption metrics for each RNA folding property that are significant for each constraint score for all SNPs

Whole Transcript	GERP	AF
Stability	100	0
Diversity	100	50
Structure	100	75

Table 1: Percentage of disruption metrics for each RNA folding property that are significant for each constraint score for all SNPs; cells in red text indicate percentages greater than or equal to 50% and blue highlighted cells indicate percentages greater than or equal to 75%

4.2. 5’ Untranslated Region (5’ UTR)

Figure 4/Table 2 show that correlations between GERP++ scores and RNA folding disruptions were significant for a moderate amount of disruption metrics in all three RNA folding properties when analyzing 5’ UTR SNPs (stability, 62.5%; diversity,

50%; structure 62.5%). Correlations between population allele frequencies and RNA folding disruptions were significant for only some structure (50%) disruption metrics.

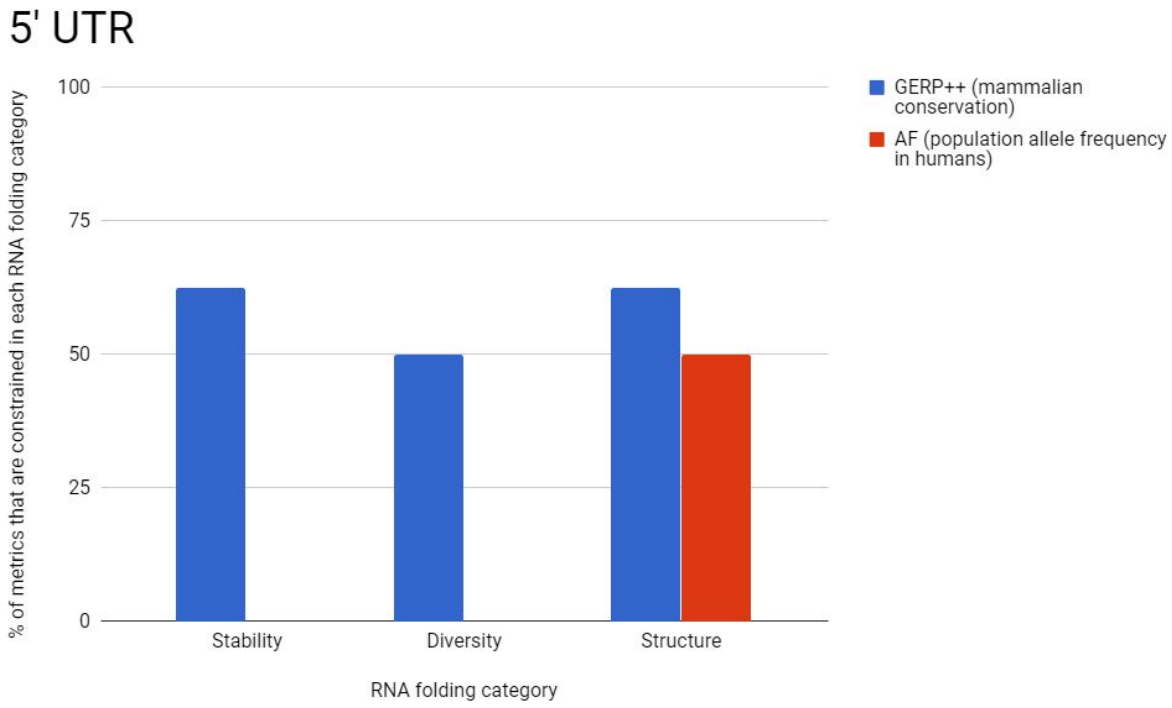


Figure 4: Percentage of disruption metrics for each RNA folding property that are significant for each constraint score for 5' UTR SNPs

5' UTR	GERP	AF
Stability	62.5	0
Diversity	50	0
Structure	62.5	50

Table 2: Percentage of disruption metrics for each RNA folding property that are significant for each constraint score for 5' UTR SNPs; cells in red text indicate percentages greater than or equal to 50% and blue highlighted cells indicate percentages greater than or equal to 75%

4.3. 3' Untranslated Region (3' UTR)

Figure 5/Table 3 show that correlations between GERP++ scores and RNA folding disruptions were significant for some disruption metrics in stability (25%) and structure (100%) disruption metrics when analyzing 3' UTR SNPs. Correlations between population allele frequencies and RNA folding disruptions were significant for almost all metrics in all three RNA folding properties (stability, 75%; diversity, 100%, structure 100%).

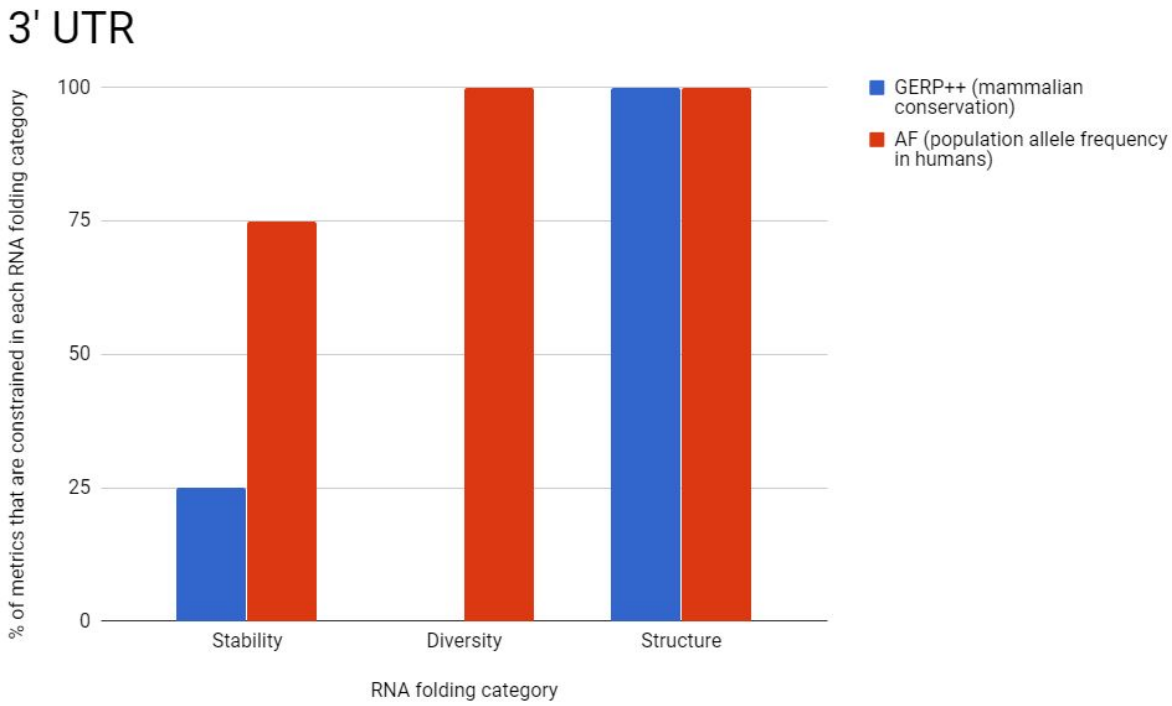


Figure 5: Percentage of disruption metrics for each RNA folding property that are significant for each constraint score for 3' UTR SNPs

3' UTR	GERP	AF
Stability	25	75
Diversity	0	100
Structure	100	100

Table 3: Percentage of disruption metrics for each RNA folding property that are significant for each constraint score for 3' UTR SNPs; cells in red text indicate percentages greater than or equal to 50% and blue highlighted cells indicate percentages

greater than or equal to 75%

4.4. Synonymous

Figure 6/Table 4 show that correlations between GERP++ scores and RNA folding disruptions were significant for none of the disruption metrics when analyzing synonymous SNPs. Correlations between population allele frequencies and RNA folding disruptions were significant for all metrics in all three RNA folding properties.

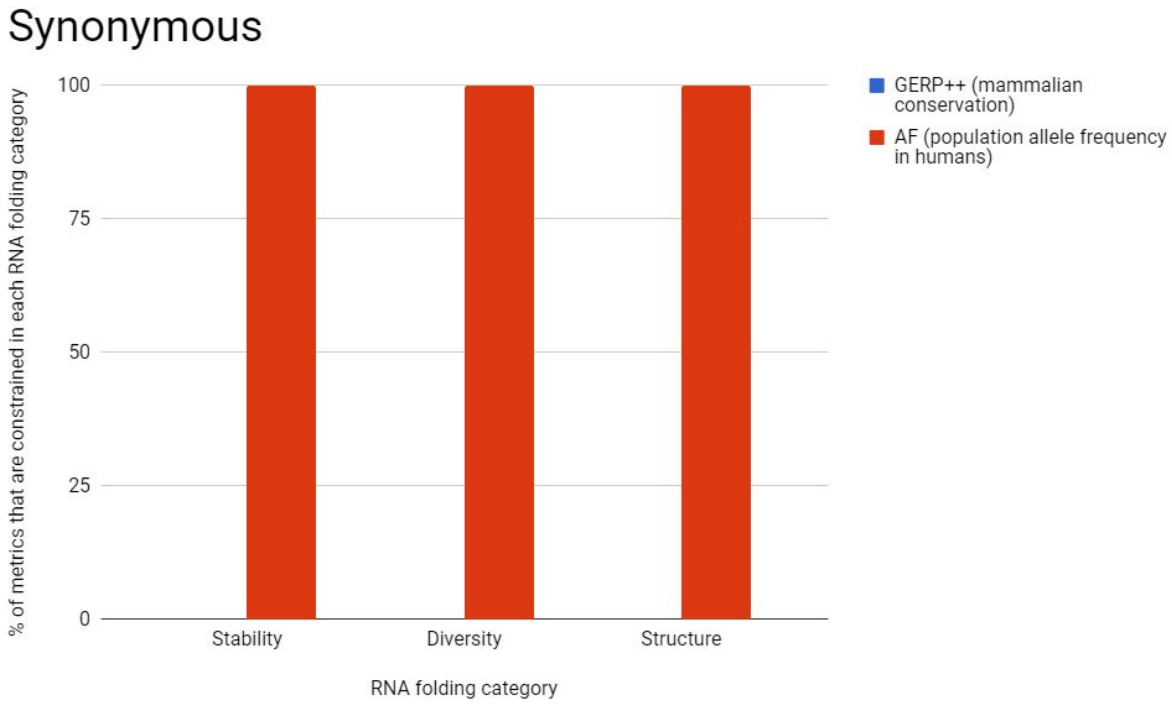


Figure 6: Percentage of disruption metrics for each RNA folding property that are significant for each constraint score for synonymous SNPs

Synonymous	GERP	AF
Stability	0	100
Diversity	0	100
Structure	0	100

Table 4: Percentage of disruption metrics for each RNA folding property that are

significant for each constraint score for synonymous SNPs; cells in red text indicate percentages greater than or equal to 50% and blue highlighted cells indicate percentages greater than or equal to 75%

4.5. Missense

Figure 7/Table 5 show that correlations between GERP++ scores and RNA folding disruptions were significant for all diversity and structure metrics and 25% of stability disruption metrics when analyzing missense SNPs. Correlations between population allele frequencies and RNA folding disruptions were also significant for all diversity and structure metrics and 25% of stability disruption metrics.

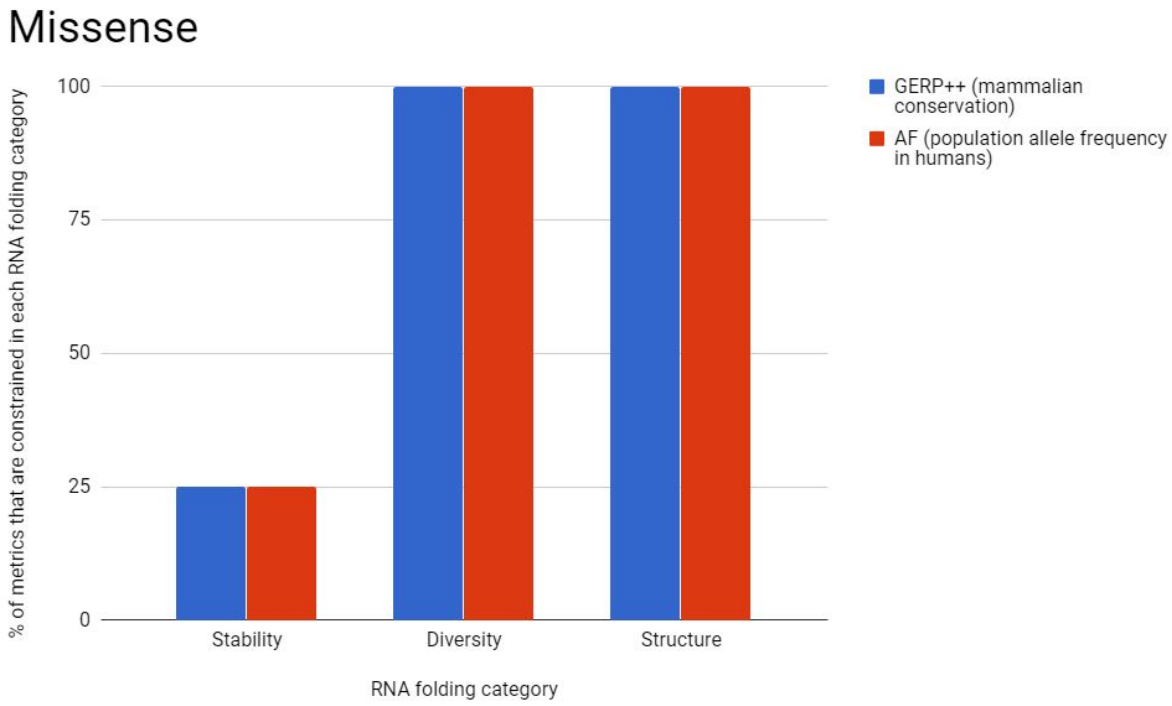


Figure 7: Percentage of disruption metrics for each RNA folding property that are significant for each constraint score for missense SNPs

Missense	GERP	AF
Stability	25	25
Diversity	100	100
Structure	100	100

Table 5: Percentage of disruption metrics for each RNA folding property that are significant for each constraint score for missense SNPs; cells in red text indicate percentages greater than or equal to 50% and blue highlighted cells indicate percentages greater than or equal to 75%

4.6 SnpRFC: "SNP mRNA Folding Consequences in Humans"

Figure 8 below shows the user interface for our app SnpRFC which quickly retrieves RNA folding statistics, disruption metric p-values, population allele frequencies, GERP++ scores, and composite disruption scores for queried SNPs.

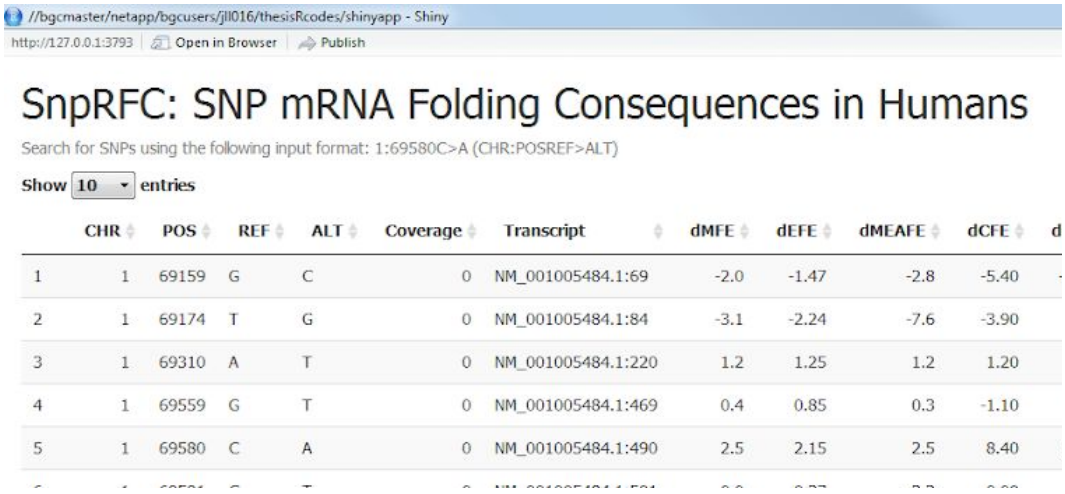


Figure 8: The user interface of "SNP mRNA Folding Consequences in Humans"

5. Discussion

Analysis of all processed SNPs as summarized in Figure 3/Table 1 showed that population allele frequency is modestly correlated with SNP-induced disruptions to RNA structure and diversity; this finding suggests that there is moderate negative selection and constraint against SNPs that are highly disruptive to RNA structure and diversity within humans. Additionally, this analysis of all processed SNPs showed that there were significant correlations between all of the RNA disruption metrics and GERP++ scores, which directly indicates that among mammalian species, the amount of base substitutions at sites with the potential to cause large RNA folding disruptions are significantly lower than the predicted mutation rate in neutral regions. These GERP++ score vs. disruption metric correlations suggest that variants that cause large RNA disruptions are selected against within mammals and that RNA folding may constrain sequence evolution for a number of mammals.

Though the analysis of all processed SNPs in the human transcriptome demonstrated moderate constraint against SNPs that are highly disruptive to RNA folding - specifically for structure and diversity, distilling the SNPs into their mRNA subregions showed that 3' UTR and synonymous SNPs had nearly all their disruption metrics correlate significantly with population allele frequencies. On the contrary, 5' UTR and missense SNPs had barely any correlations between stability metrics and population allele frequencies; moreover, 5' UTR SNPs also lacked correlations between diversity metrics and population allele frequencies. This lack of stability and diversity metric correlations in these SNPs helps explain why the overall constraint in humans against SNPs that are highly disruptive to RNA folding appears to be more moderate rather than strong.

As shown in Figures 3-7/Tables 1-5, the constraints of RNA structure disruption metrics on population allele frequency were common among SNPs from all mRNA

subregions. Past studies have shown that the position of hairpin loop structures near the 5' UTR affects translation efficiency, mRNA secondary structure within the coding region can affect or halt translation, and mRNA secondary structure in 3' UTR regions can affect microRNA binding sites which thus affects gene regulation/expression (Babendure *et al.* 2006, Chen *et al.* 2013, Fang *et al.* 2011). Thus, the common constraint on SNPs with large RNA structure disruptions may be a consequence of their effects on protein translation rates and/or appropriate folding, ultimately impacting levels of protein expression.

As shown in Figures 5&6/Tables 3&4, despite 3' UTR and synonymous SNPs having many disruption metrics significantly correlate to population allele frequencies, they did not have many metrics significantly correlate to GERP++ scores. This finding lends to the notion that selection against highly disruptive 3' UTR and synonymous SNPs may be a constraint that tends to be more specific to humans. Compounding on this human-specific notion, it must be noted that the percentage of stability metrics correlated with GERP++ scores was only above 50% in 5' UTR SNPs, while the percentage of stability metrics correlated with population allele frequency was above 50% in both 3' UTR and synonymous SNPs. We previously revealed that the plethora of correlated RNA folding metrics of both 3' UTR and synonymous SNPs to population allele frequency suggests the overall constraint of RNA folding disruptions is likely underrepresented when viewing the whole transcript results; however, the additional notion that RNA stability disruptions correlate to population allele frequencies more noticeably than to GERP++ scores suggests that RNA stability disruptions may be a more human-specific folding property.

6. Conclusions and Future Work

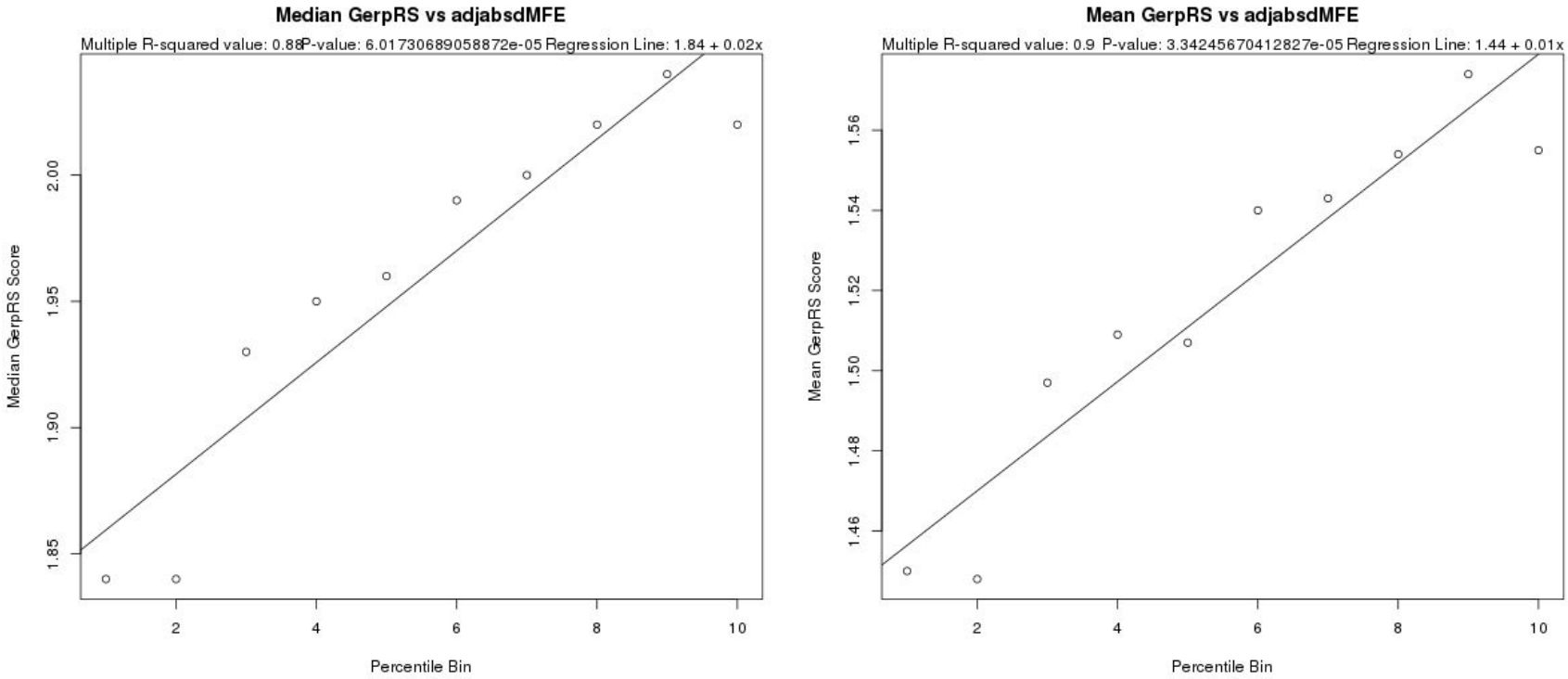
Selection against polymorphisms that cause large RNA folding disruptions appears to be present in both humans and other mammals. Our whole mRNA transcriptome analysis lends to the notion that mRNA is not purely an intermediate that contains information for protein synthesis, but rather the inherent properties of the mRNA such as the structure, stability, and diversity may influence protein production and ultimately expression. With this analysis, we were able to support our hypothesis that RNA folding plays a role in human health and disease. Moreover, we were able to create our app, SnpRFC, to annotate SNPs based on data extracted from our pipeline, and we plan to further develop SnpRFC's composite disruption scores into new metrics for RNA-based sequence variant analysis for disease. Our pipeline, coupled with SnpRFC, may be used to examine datasets such as The Cancer Genome Atlas to see if genes that are correlated with specific disease phenotypes may have certain characteristic RNA folding patterns. Lastly, while our study comprehensively folds flanking sequences derived from mRNA transcripts, there exists a pool of noncoding RNA transcripts in humans. Applying our pipeline to these noncoding RNAs may further elucidate the impacts of RNA folding on human health and disease.

7. References

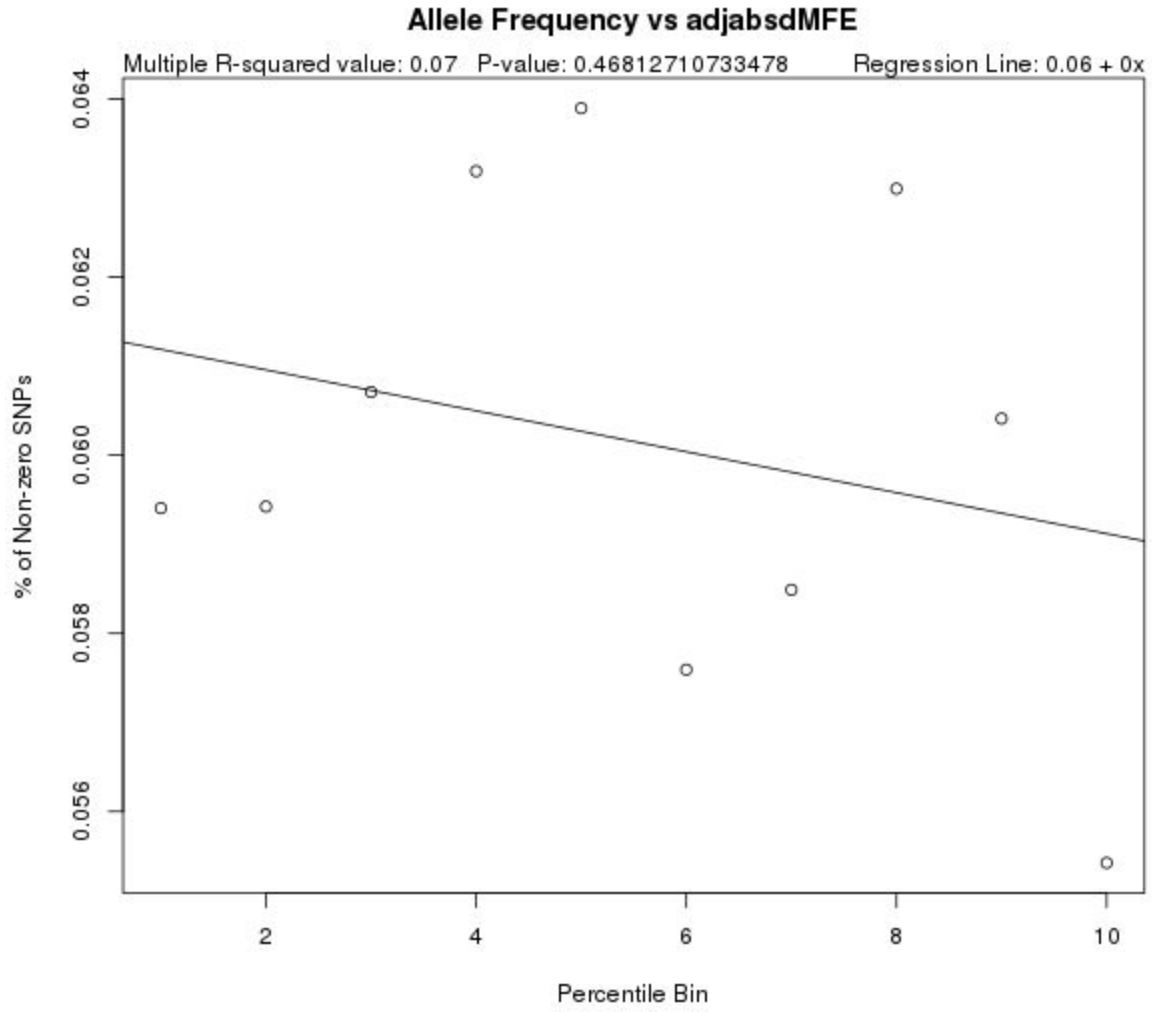
1) Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., ... & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature methods*, *7*(4), 248.
2) Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols*, *4*(7), 1073.
3) Holmila, R., Fouquet, C., Cadranel, J., Zalcman, G., & Soussi, T. (2003). Splice mutations in the p53 gene: case report and review of the literature. *Human mutation*, *21*(1), 101-102.
4) Raponi, M., & Baralle, D. (2010). Alternative splicing: good and bad effects of translationally silent substitutions. *The FEBS journal*, *277*(4), 836-840.
5) Sauna, Z. E., & Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. *Nature Reviews Genetics*, *12*(10), 683.
6) Gartner, J. J., Parker, S. C., Prickett, T. D., Dutton-Regester, K., Stitzel, M. L., Lin, J. C., ... & Gotea, V. (2013). Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. *Proceedings of the National Academy of Sciences*, *110*(33), 13481-13486.
7) Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., & Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. *Cell*, *156*(6), 1324-1335.
8) Gotea, V., Gartner, J. J., Qutob, N., Elnitski, L., & Samuels, Y. (2015). The functional relevance of somatic synonymous mutations in melanoma and other cancers. *Pigment cell & melanoma research*, *28*(6), 673-684.
9) Soussi, T., Taschner, P. E., & Samuels, Y. (2017). Synonymous Somatic Variants in Human Cancer Are Not Infamous: A Plea for Full Disclosure in Databases and Publications. *Human mutation*, *38*(4), 339-342.
10) Johnson, A. D., Trumbower, H., & Sadee, W. (2011). RNA structures affected by single nucleotide polymorphisms in transcribed regions of the human genome.
11) Vilmi, T., Moilanen, J. S., Finnilä, S., & Majamaa, K. (2005). Sequence variation in the tRNA genes of human mitochondrial DNA. *Journal of molecular evolution*, *60*(5), 587-597.
12) Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, *6*(1), 26.
13) Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R., & Hofacker, I. L. (2008). The vienna RNA websuite. *Nucleic acids research*, *36*(suppl_2), W70-W74.

14) Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., & Haussler, D. (2003). Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences*, *100*(20), 11484-11489.
15) Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., ... & Tukiainen, T. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285.
16) Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., ... & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80-92.
17) Cooper, G. M., Stone, E. A., Asimenos, G., Green, E. D., Batzoglou, S., & Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome research*, *15*(7), 901-913.
18) Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS computational biology*, *6*(12), e1001025.
19) Hamasaki‐Katagiri, N., Lin, B. C., Simon, J., Hunt, R. C., Schiller, T., Russek‐Cohen, E., ... & Kimchi‐Sarfaty, C. (2017). The importance of mRNA structure in determining the pathogenicity of synonymous and non‐synonymous mutations in haemophilia. *Haemophilia*, *23*(1).
20) Babendure, J. R., Babendure, J. L., Ding, J. H., & Tsien, R. Y. (2006). Control of mammalian translation by mRNA structure near caps. *Rna*, *12*(5), 851-861.
21) Chen, C., Zhang, H., Broitman, S. L., Reiche, M., Farrell, I., Cooperman, B. S., & Goldman, Y. E. (2013). Dynamics of translation by single ribosomes through mRNA secondary structures. *Nature Structural and Molecular Biology*, *20*(5), 582.
22) Fang, Z., & Rajewsky, N. (2011). The impact of miRNA target sites in coding sequences and in 3′ UTRs. *PloS one*, *6*(3), e18067.

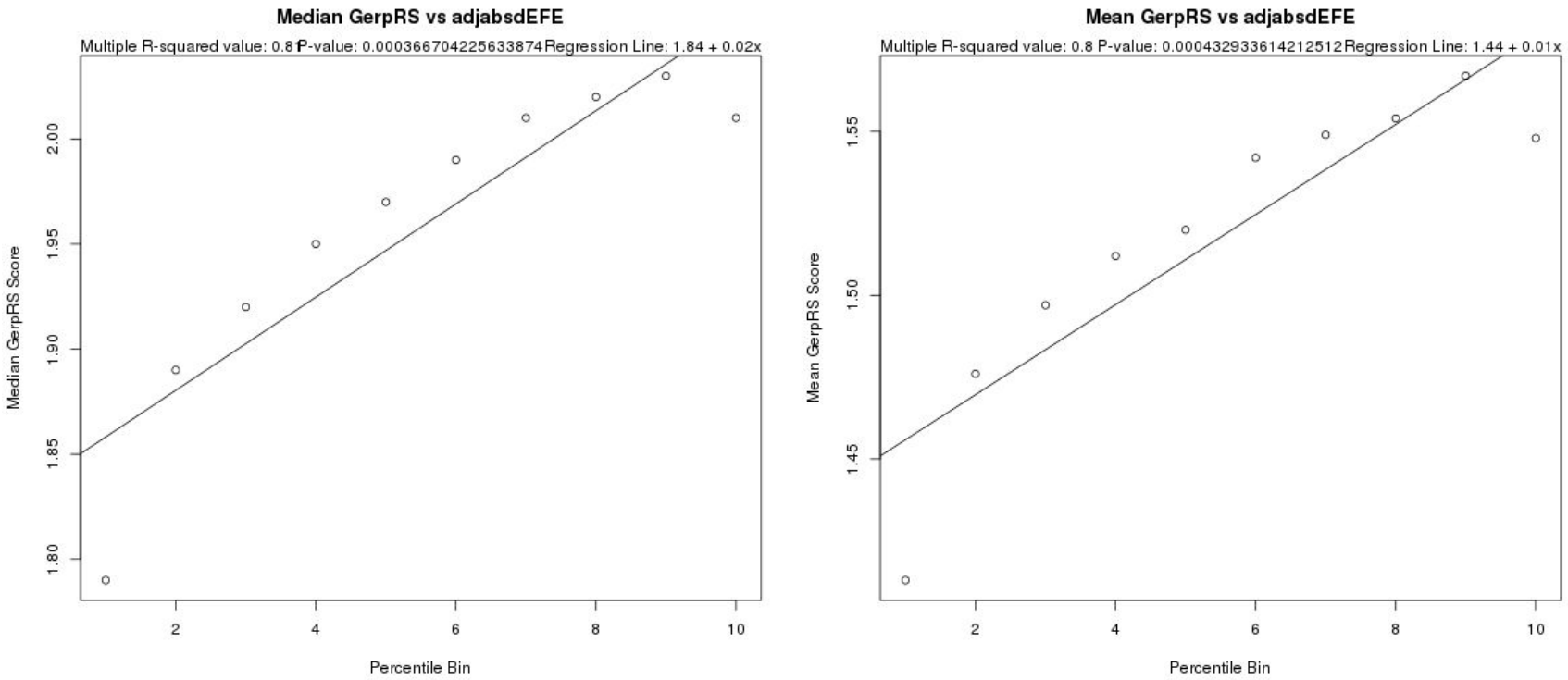
8. Supplementary Figures



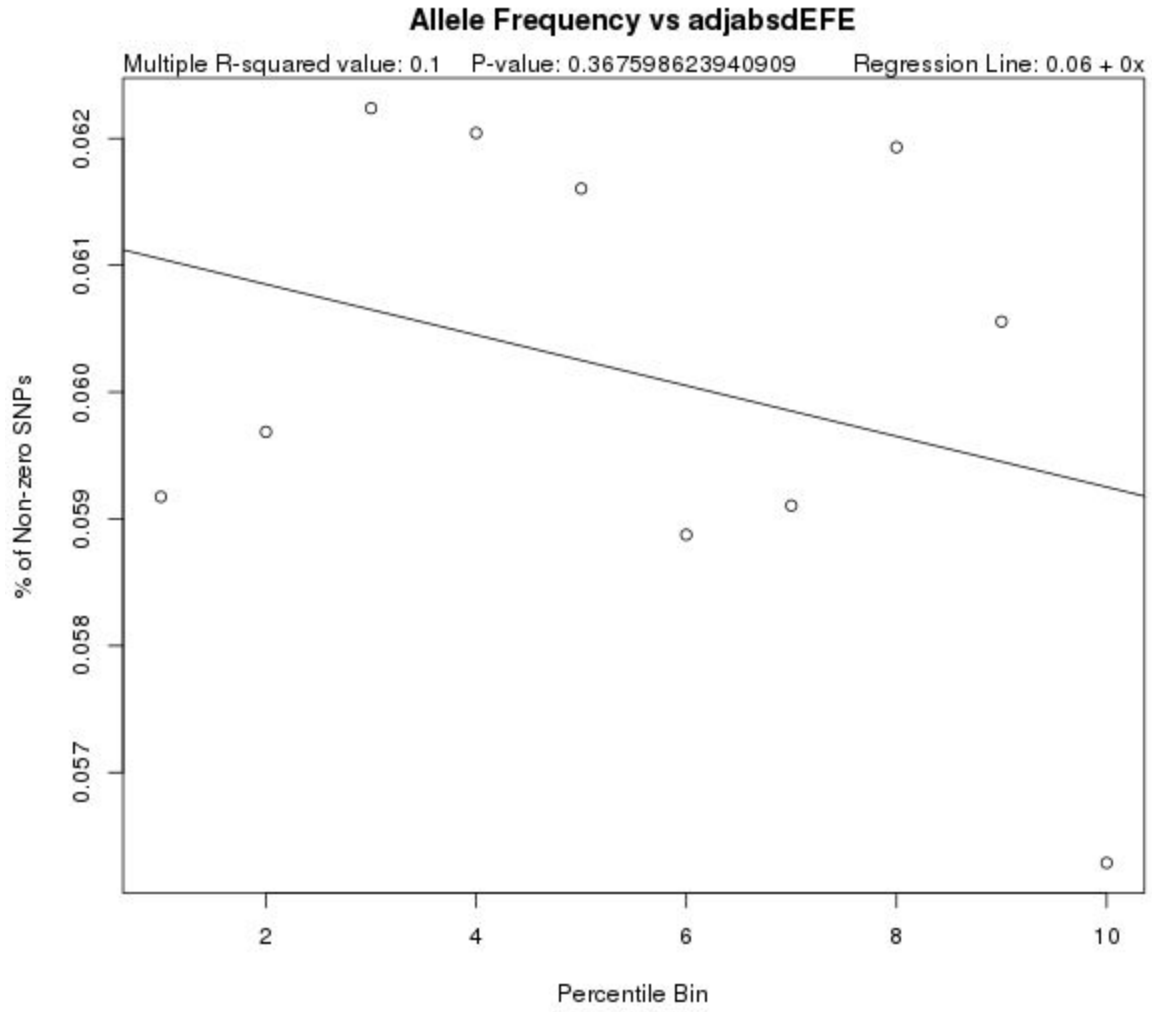
Supplementary Figure 1: Mean/Median GERP Score vs. Binned Change in Minimum Free Energy (dMFE) for All Transcript Regions



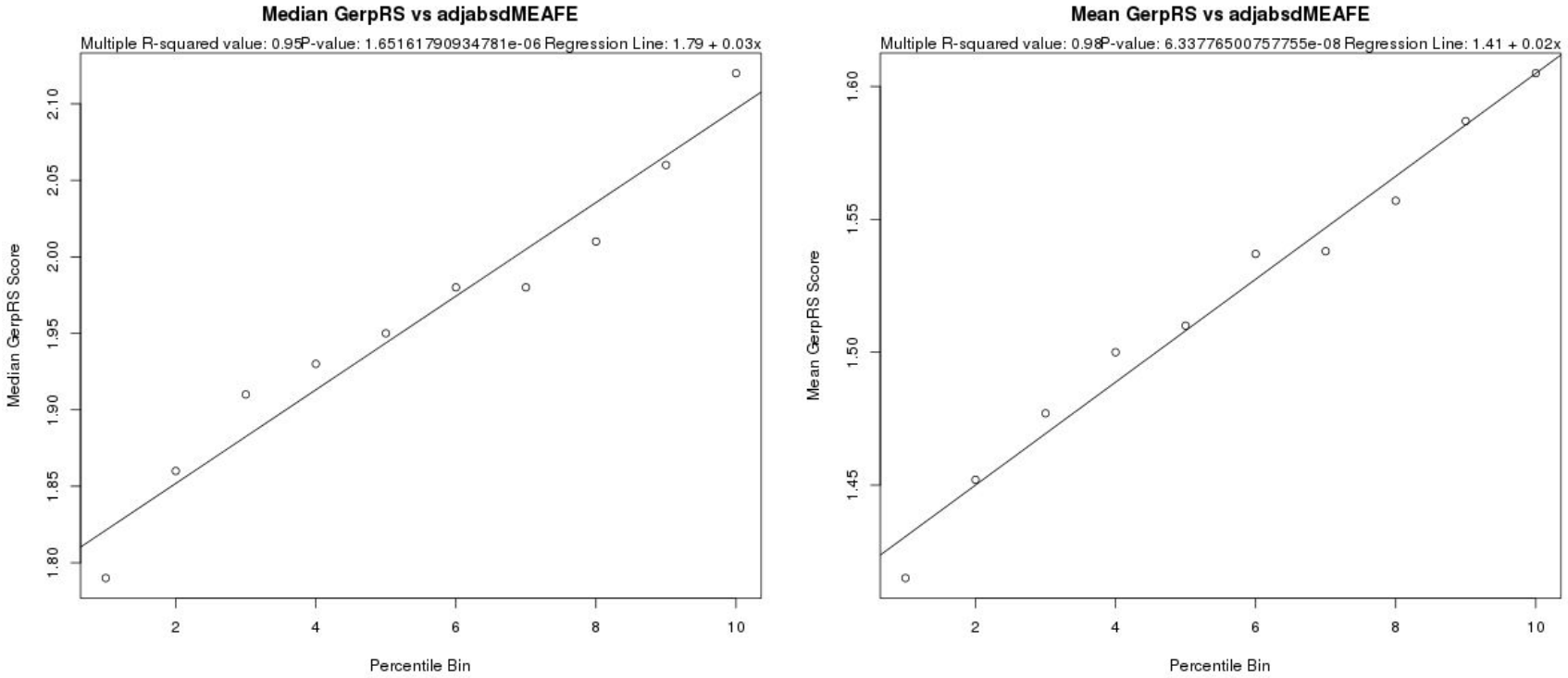
Supplementary Figure 2: % Non-zero Allele Frequency vs. Binned Change in Minimum Free Energy (dMFE) for All Transcript Regions



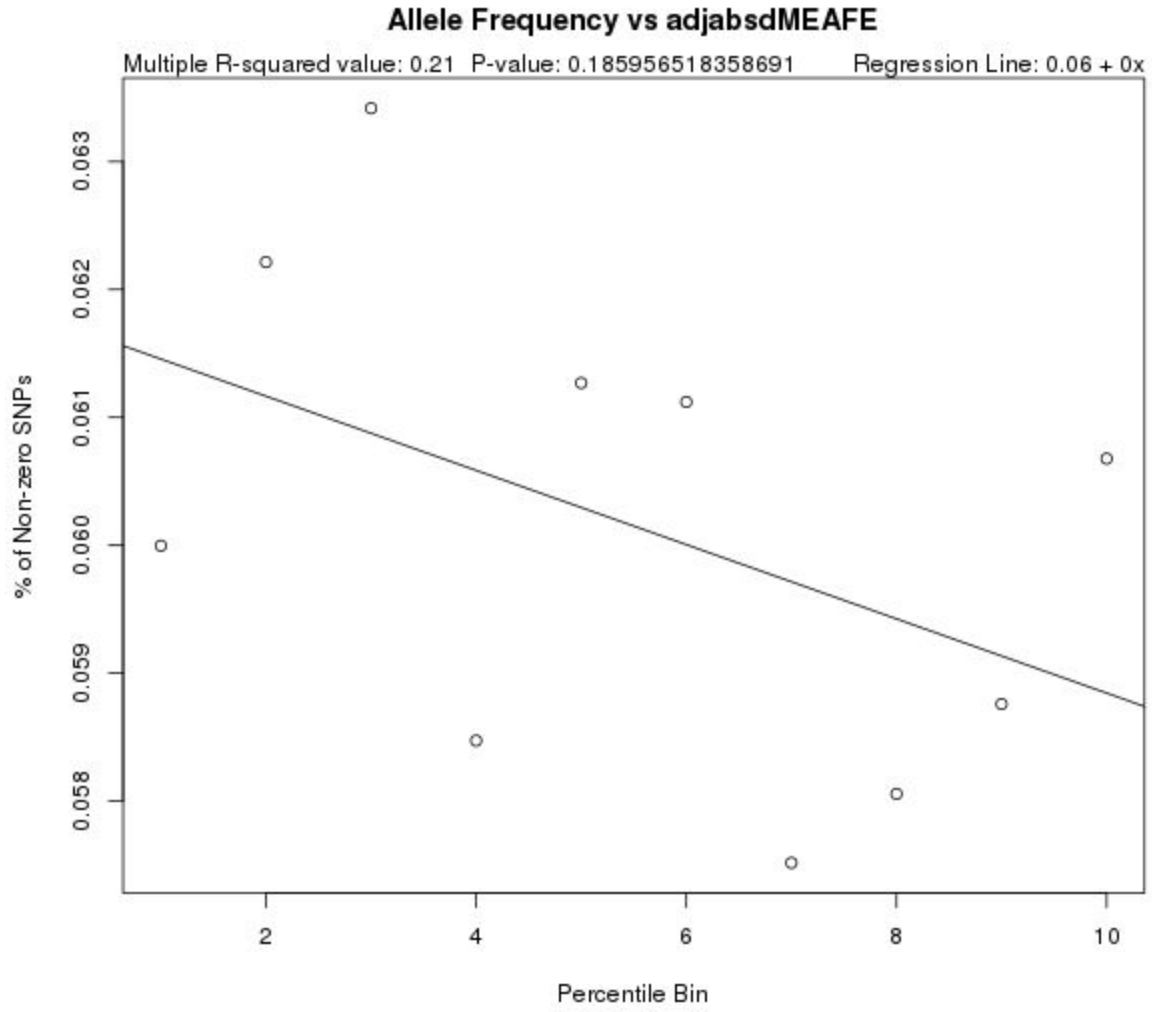
Supplementary Figure 3: Mean/Median GERP Score vs. Binned Change in Ensemble Free Energy (dEFE) for All Transcript Regions



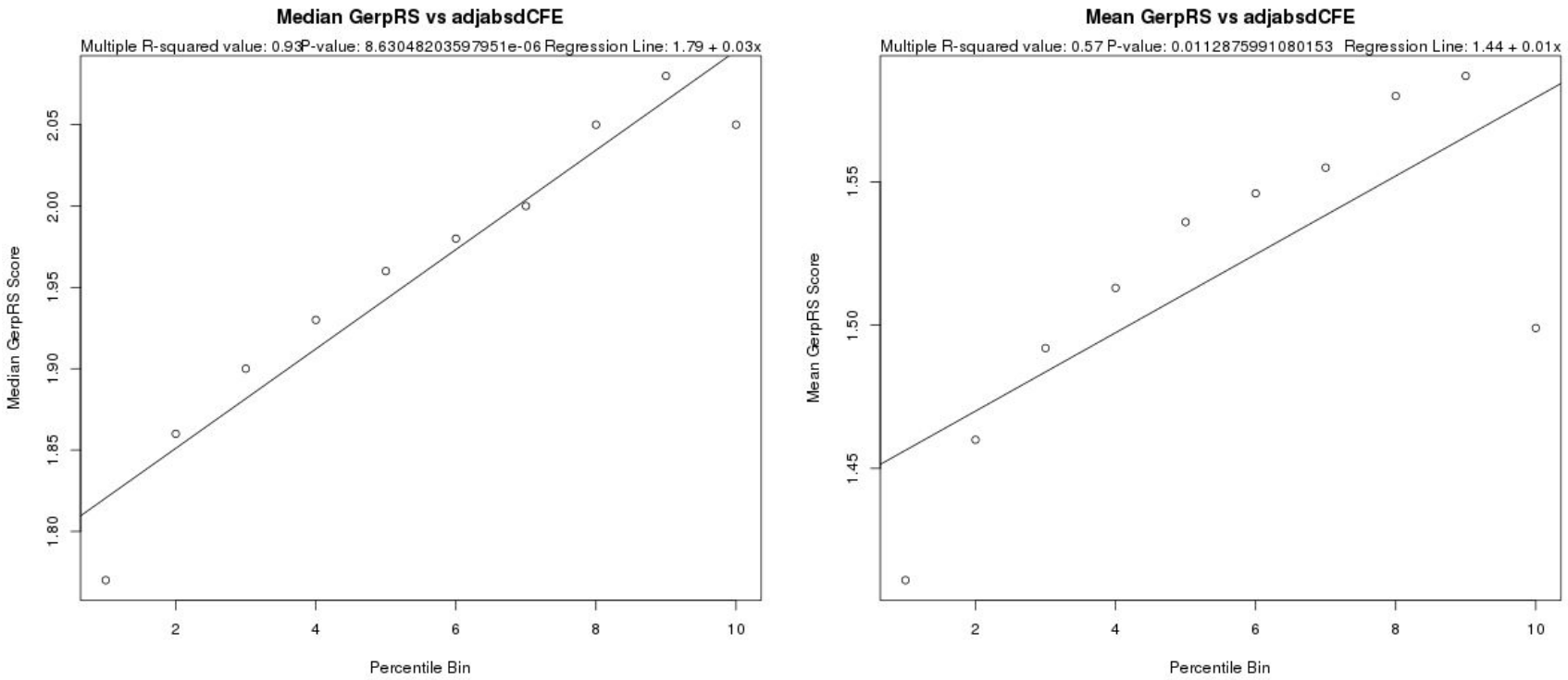
Supplementary Figure 4: % Non-zero Allele Frequency vs. Binned Change in Ensemble Free Energy (dEFE) for All Transcript Regions



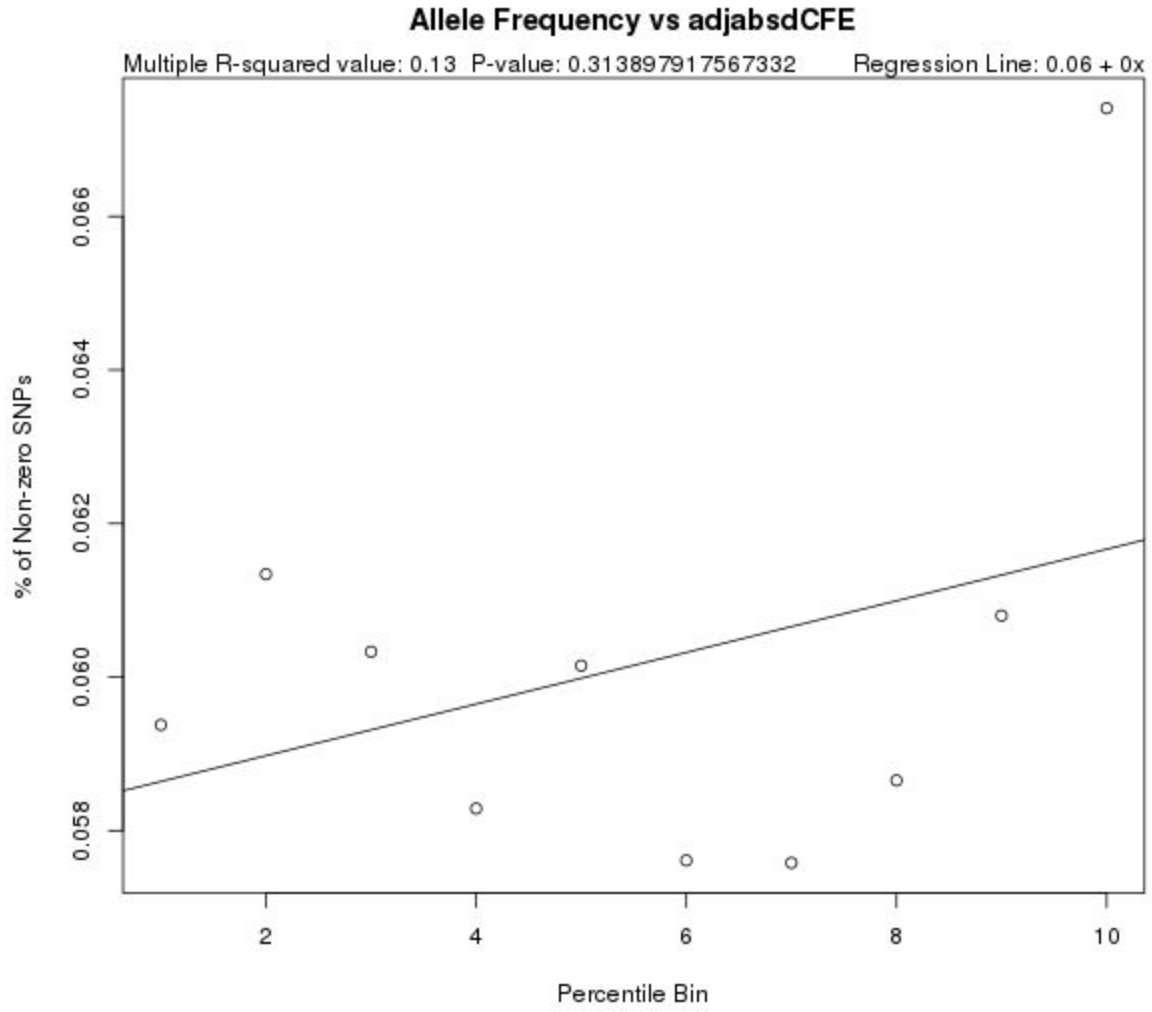
Supplementary Figure 5: Mean/Median GERP Score vs. Binned Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for All Transcript Regions



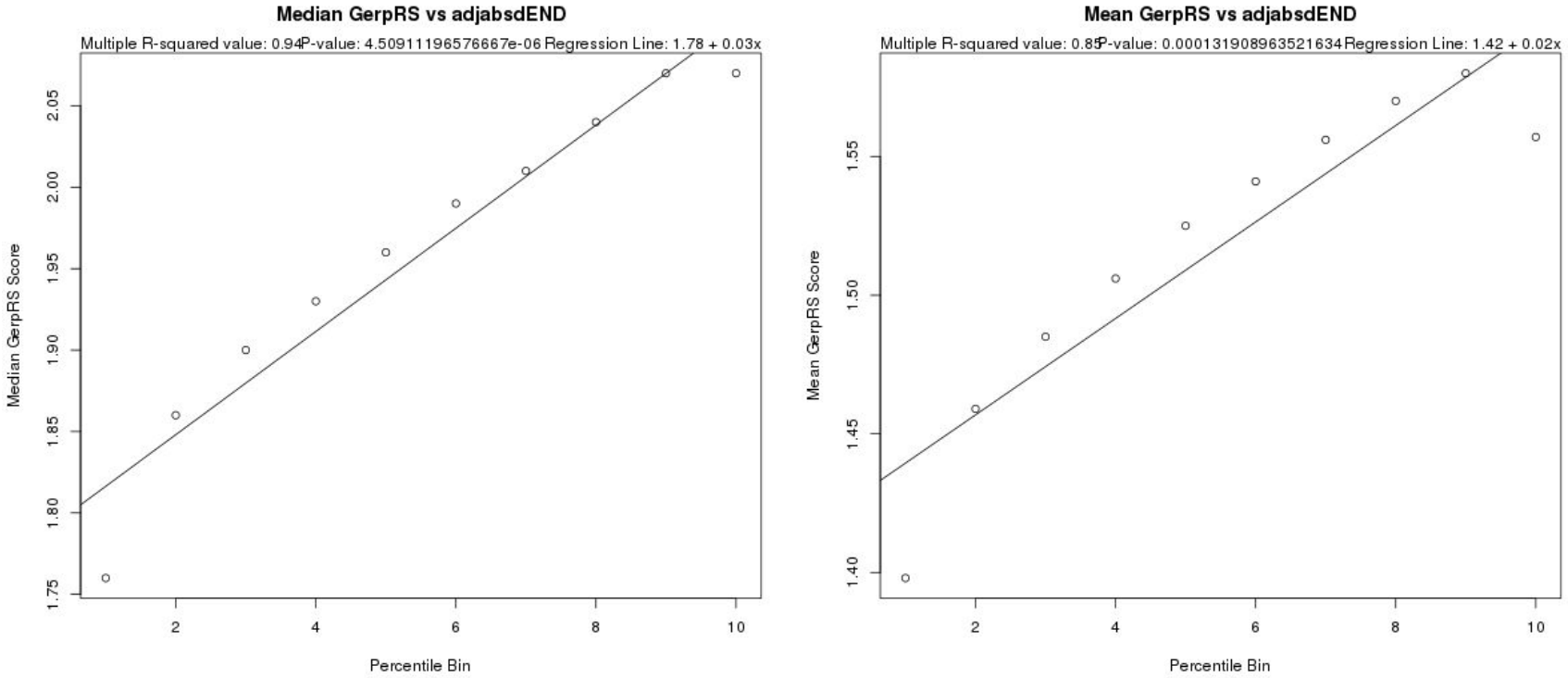
Supplementary Figure 6: % Non-zero Allele Frequency vs. Binned Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for All Transcript Regions



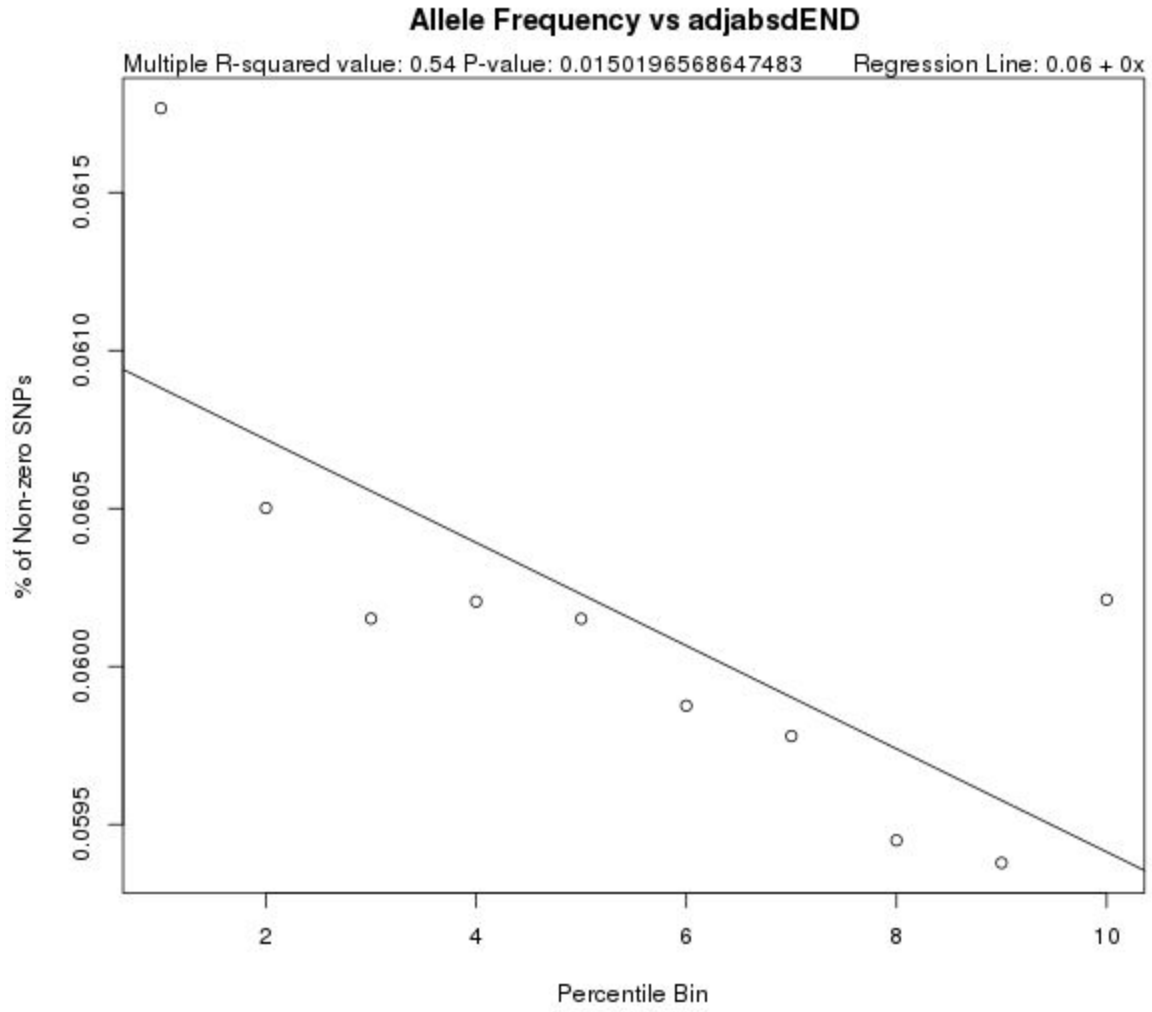
Supplementary Figure 7: Mean/Median GERP Score vs. Binned Change in Free Energy of the Centroid (dCFE) for All Transcript Regions



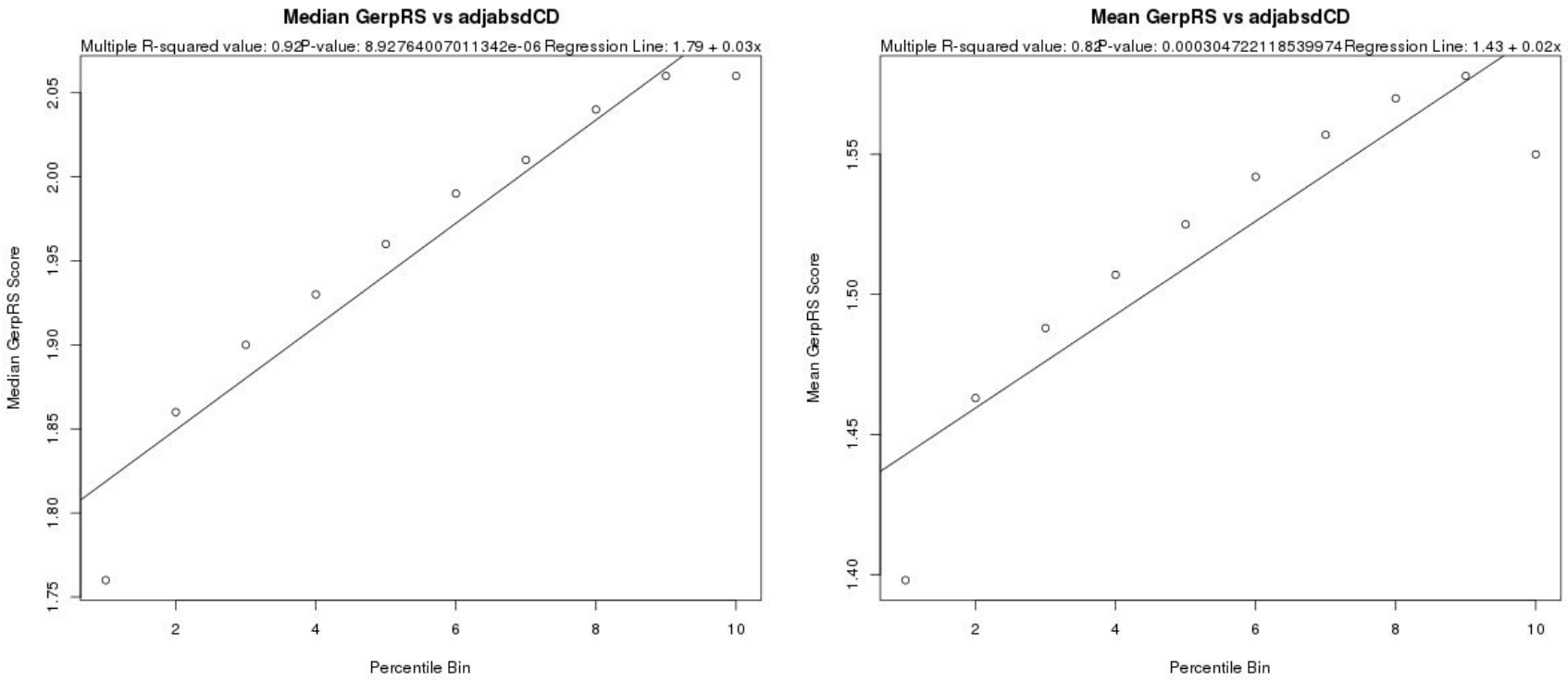
Supplementary Figure 8: % Non-zero Allele Frequency vs. Binned Change in Free Energy of the Centroid (dCFE) for All Transcript Regions



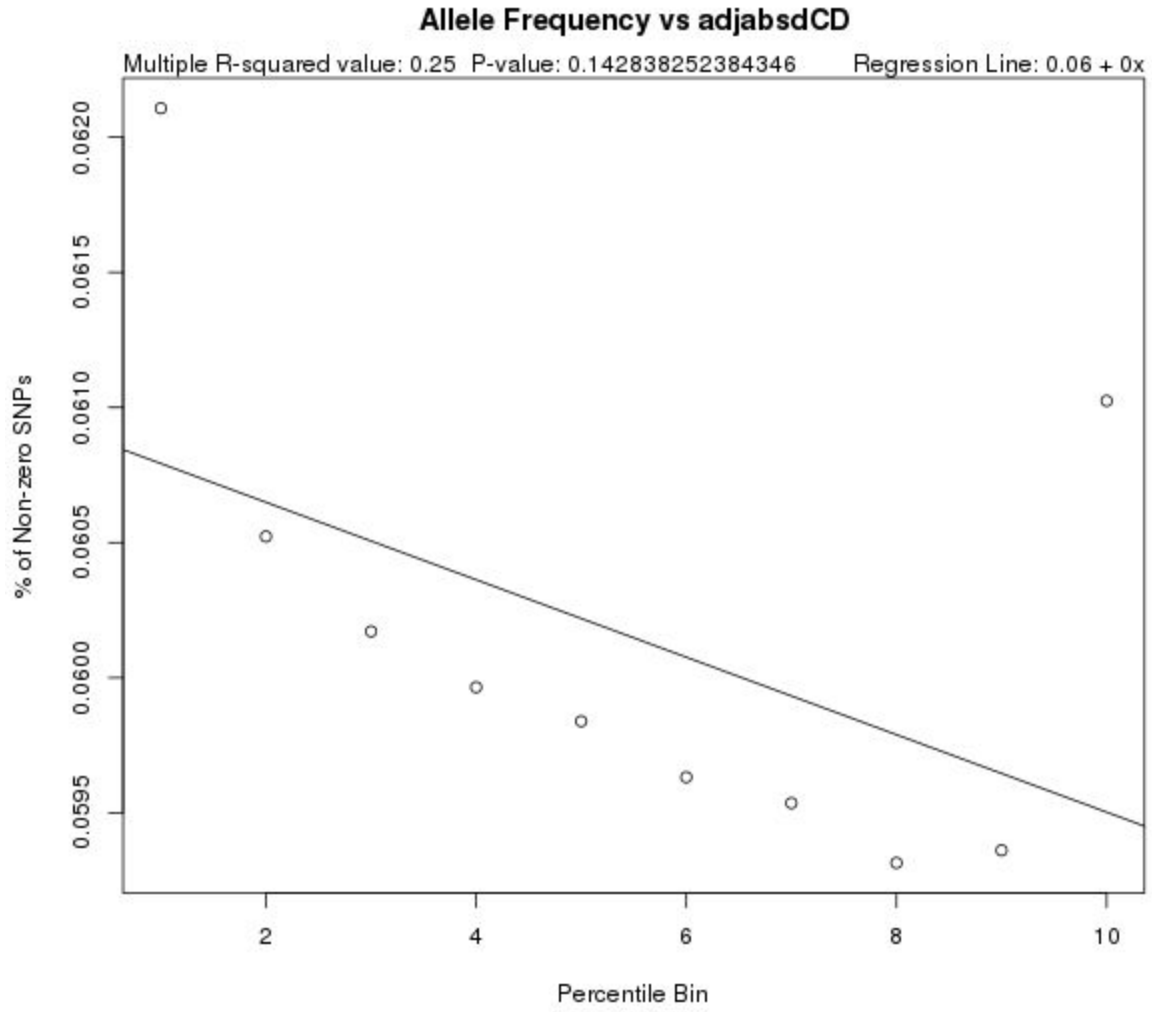
Supplementary Figure 9: Mean/Median GERP Score vs. Binned Change in Ensemble Diversity (dEND) for All Transcript Regions



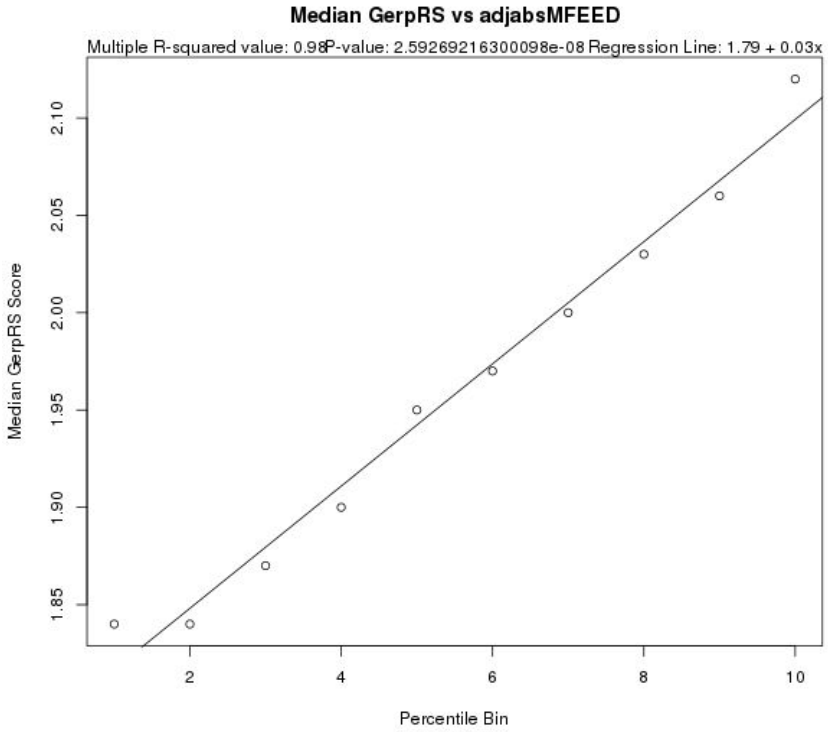
Supplementary Figure 10: % Non-zero Allele Frequency vs. Binned Change in Ensemble Diversity (dEND) for All Transcript Regions

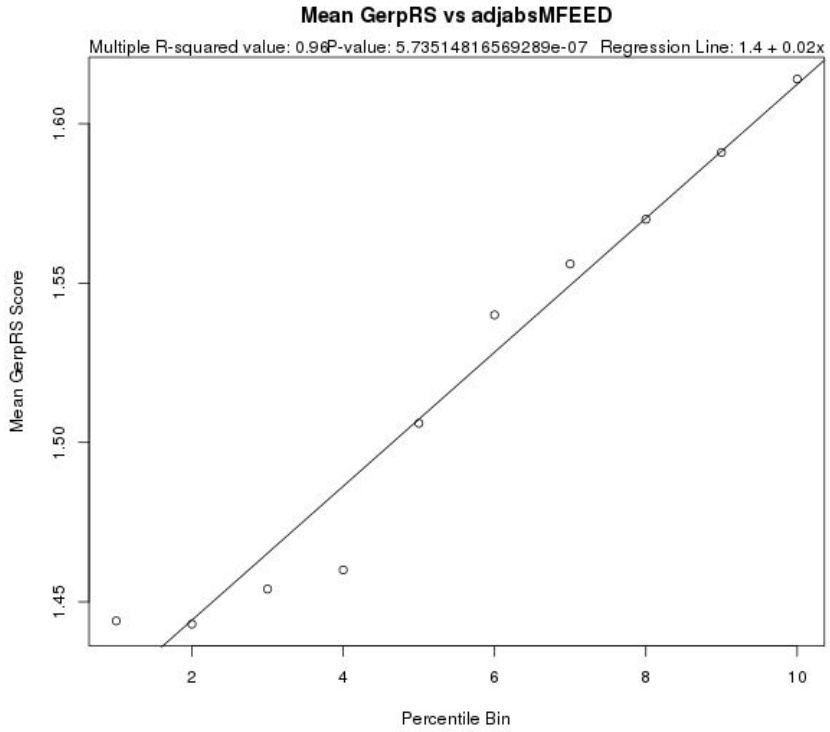


Supplementary Figure 11: Mean/Median GERP Score vs. Binned Change in Distance of the Ensemble of Structures to the Centroid (dCD) for All Transcript Regions

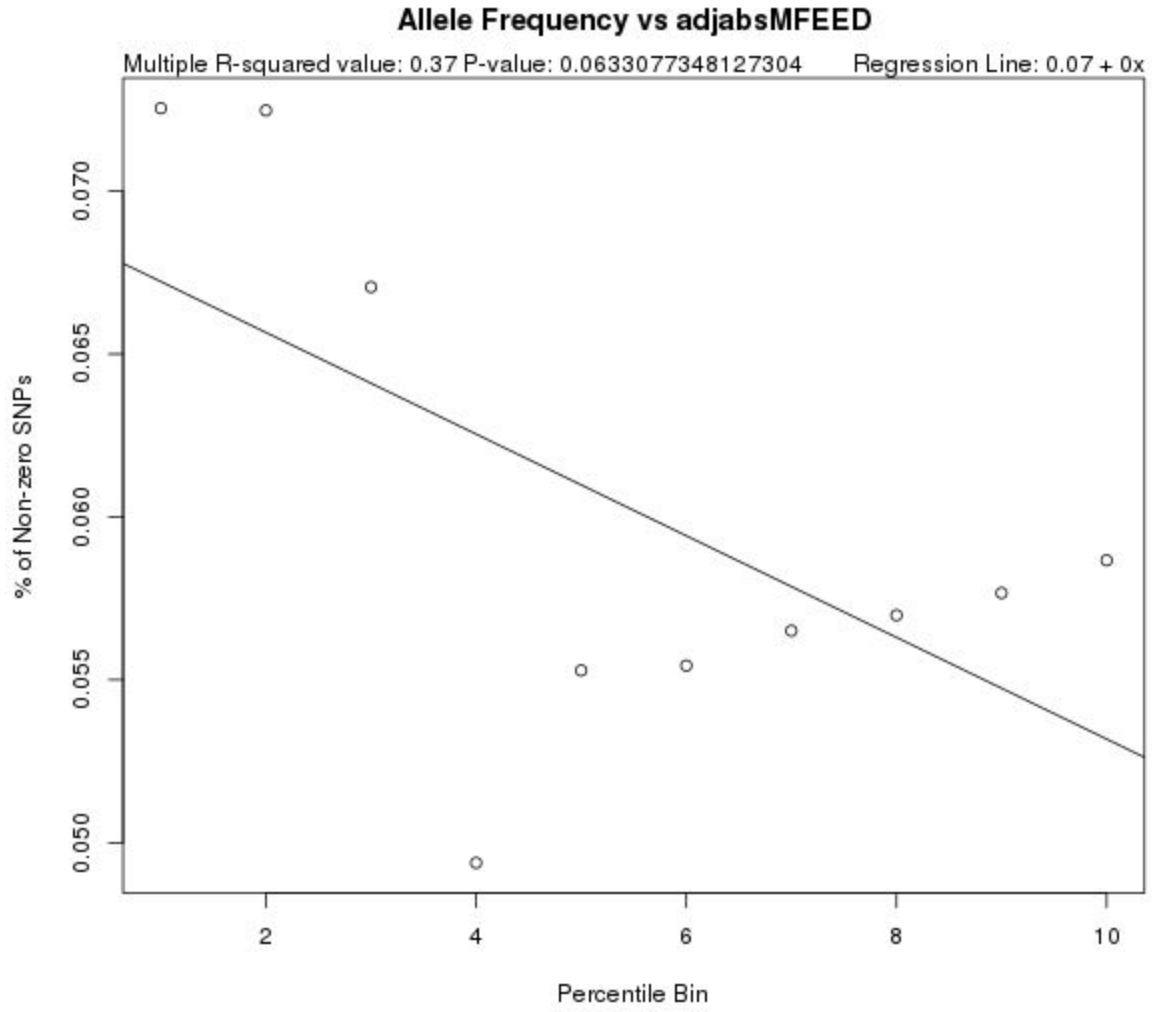


Supplementary Figure 12: % Non-zero Allele Frequency vs. Binned Change in Distance of the Ensemble of Structures to the Centroid (dCD) for All Transcript Regions

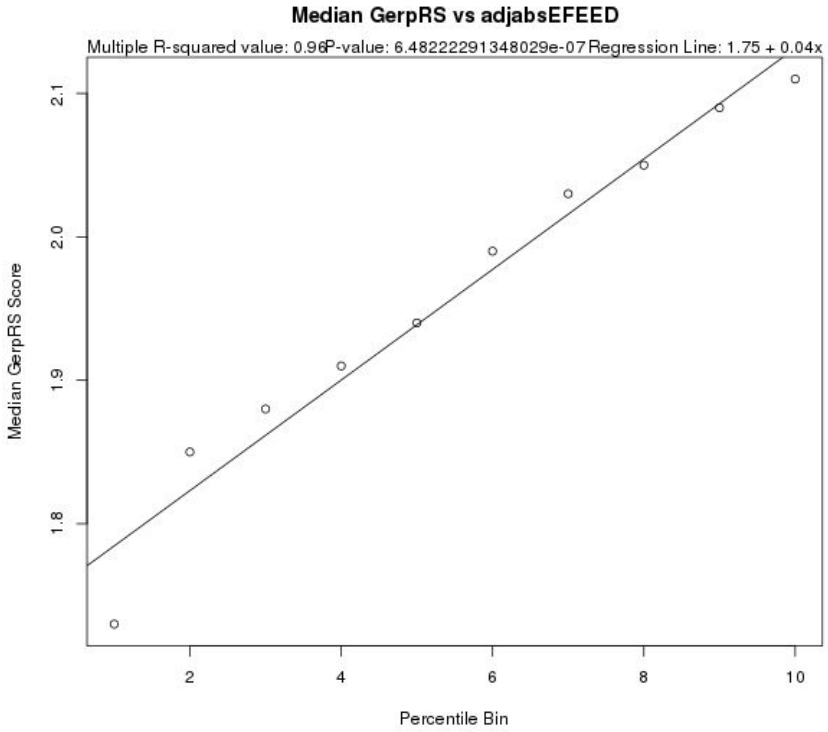


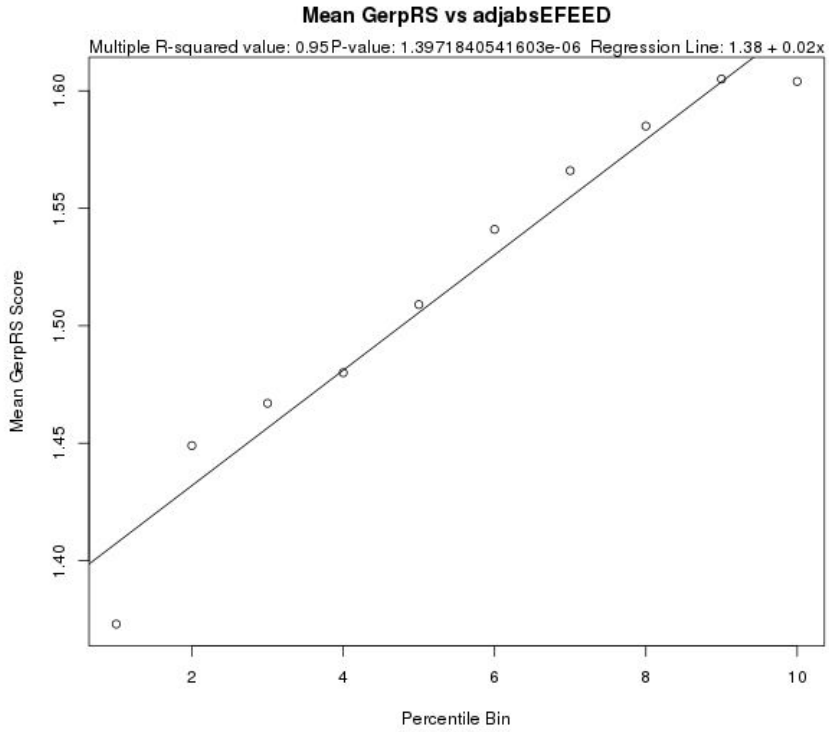


Supplementary Figure 13: Mean/Median GERP Score vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for All Transcript Regions

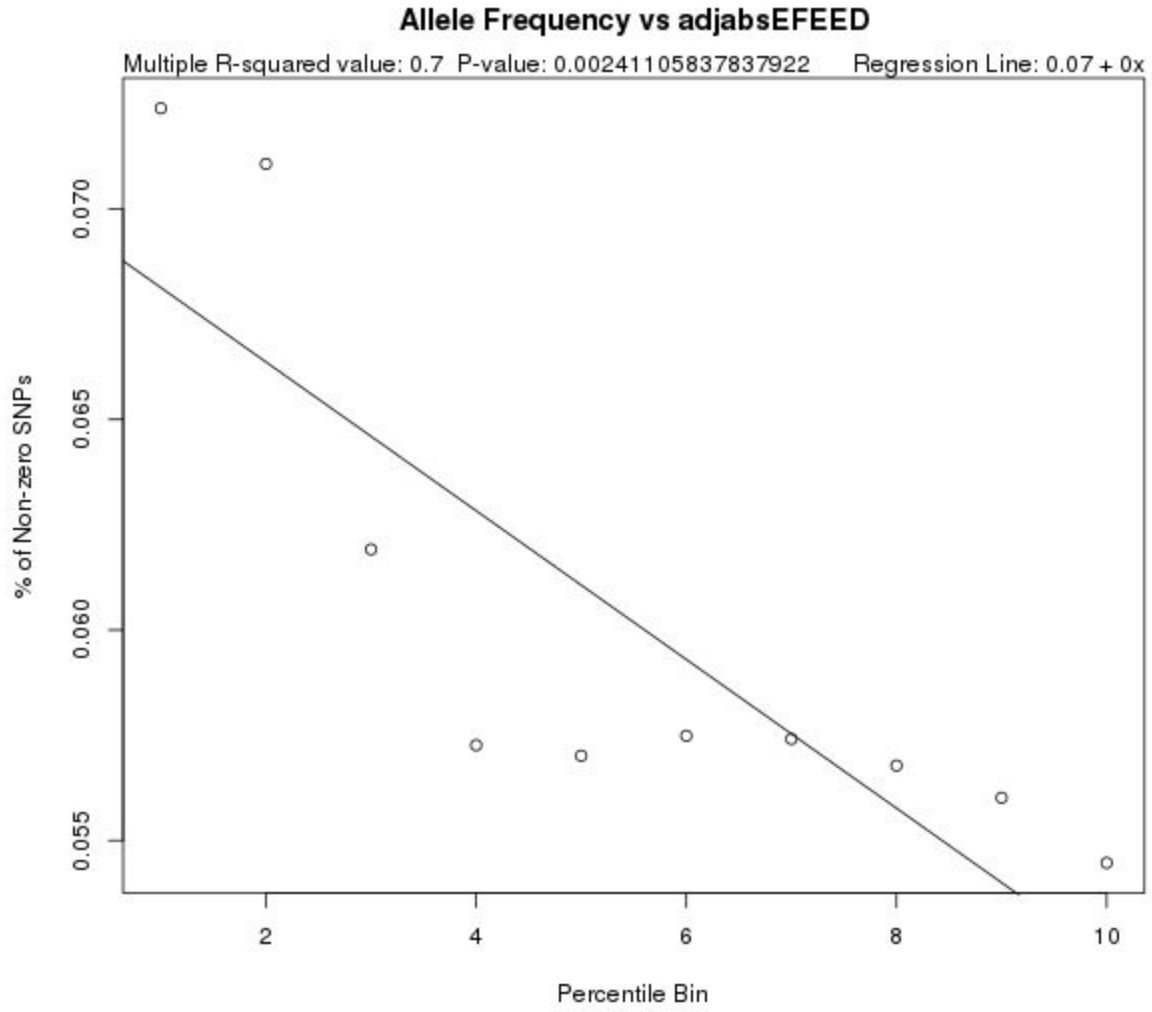


Supplementary Figure 14: % Non-zero Allele Frequency vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for All Transcript Regions

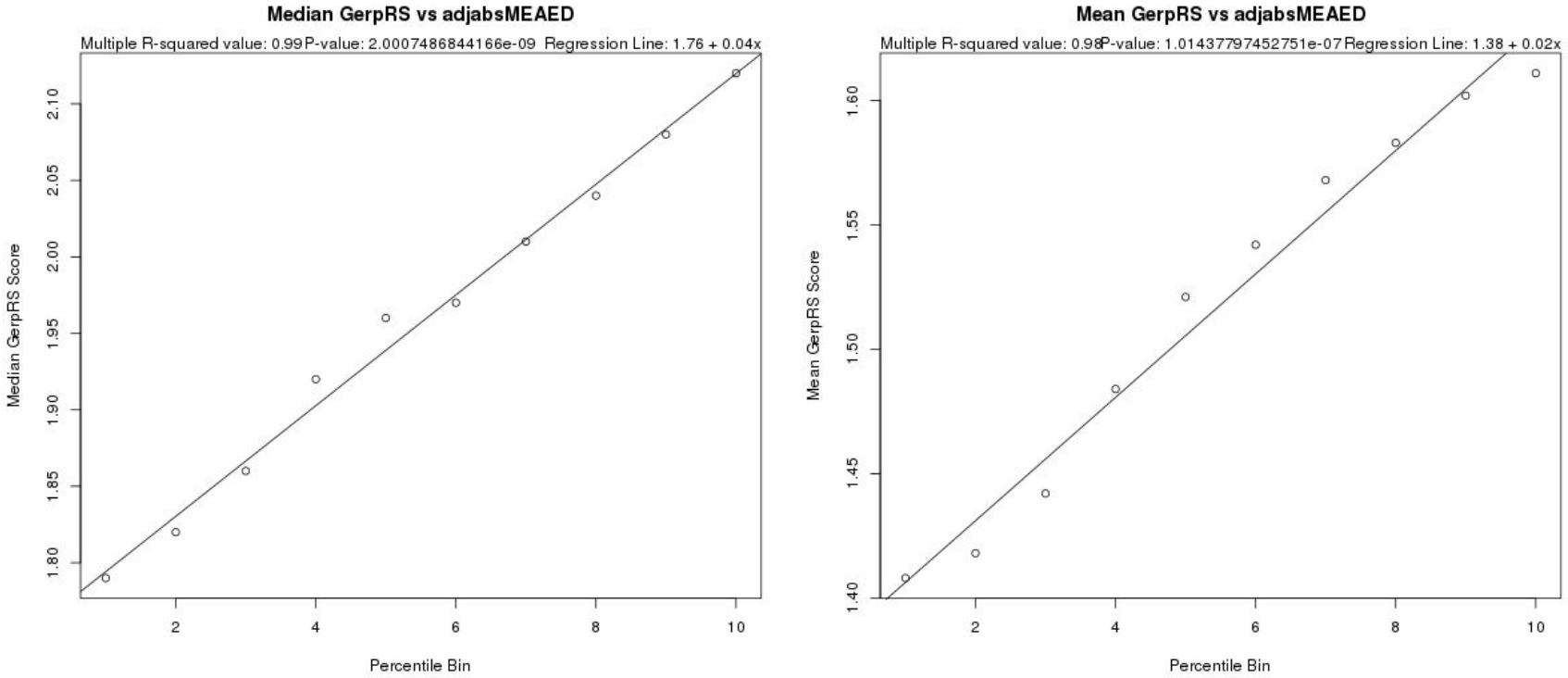




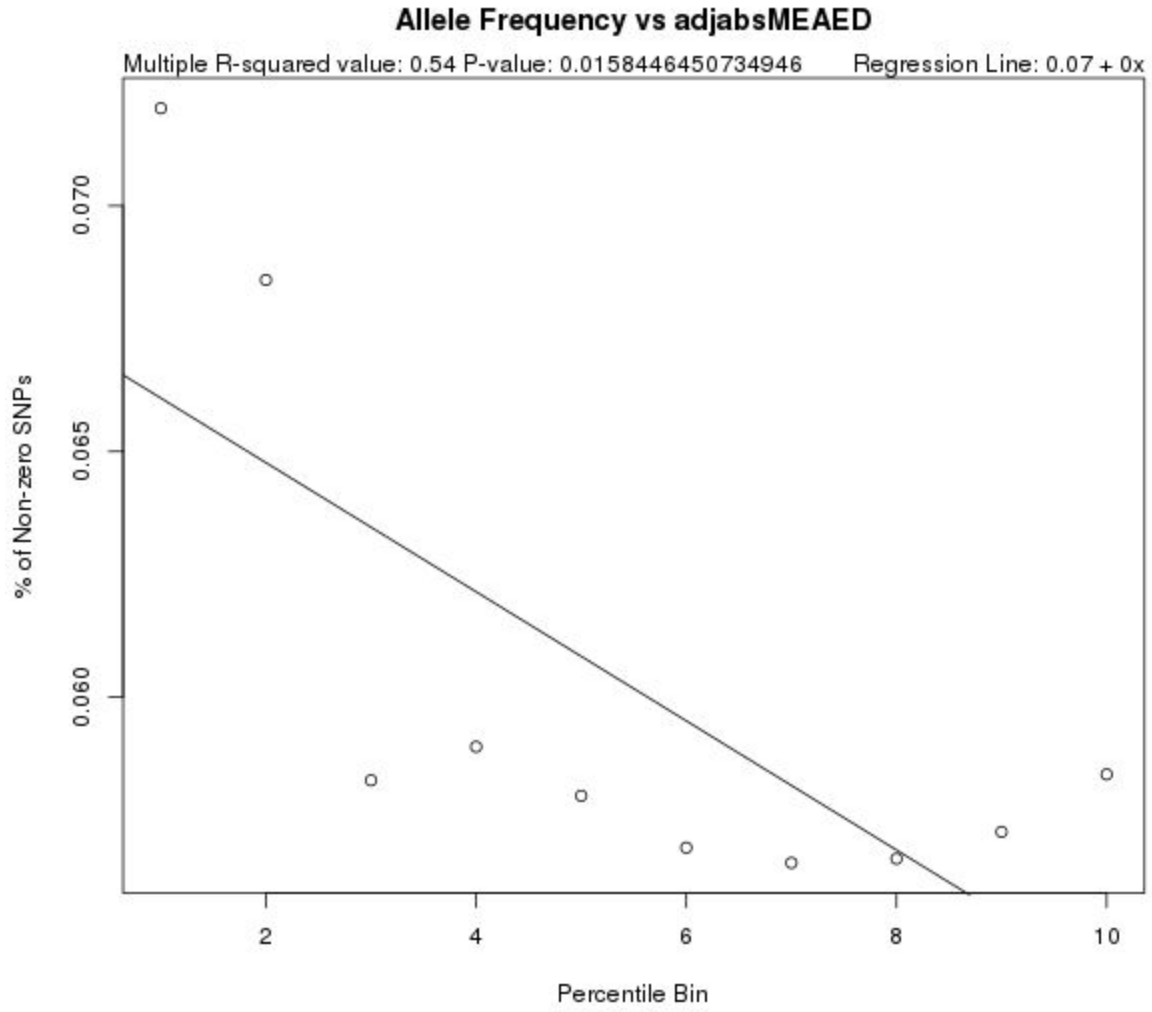
Supplementary Figure 15: Mean/Median GERP Score vs. Edit Distance Between Ensembles (EFEED) for All Transcript Regions



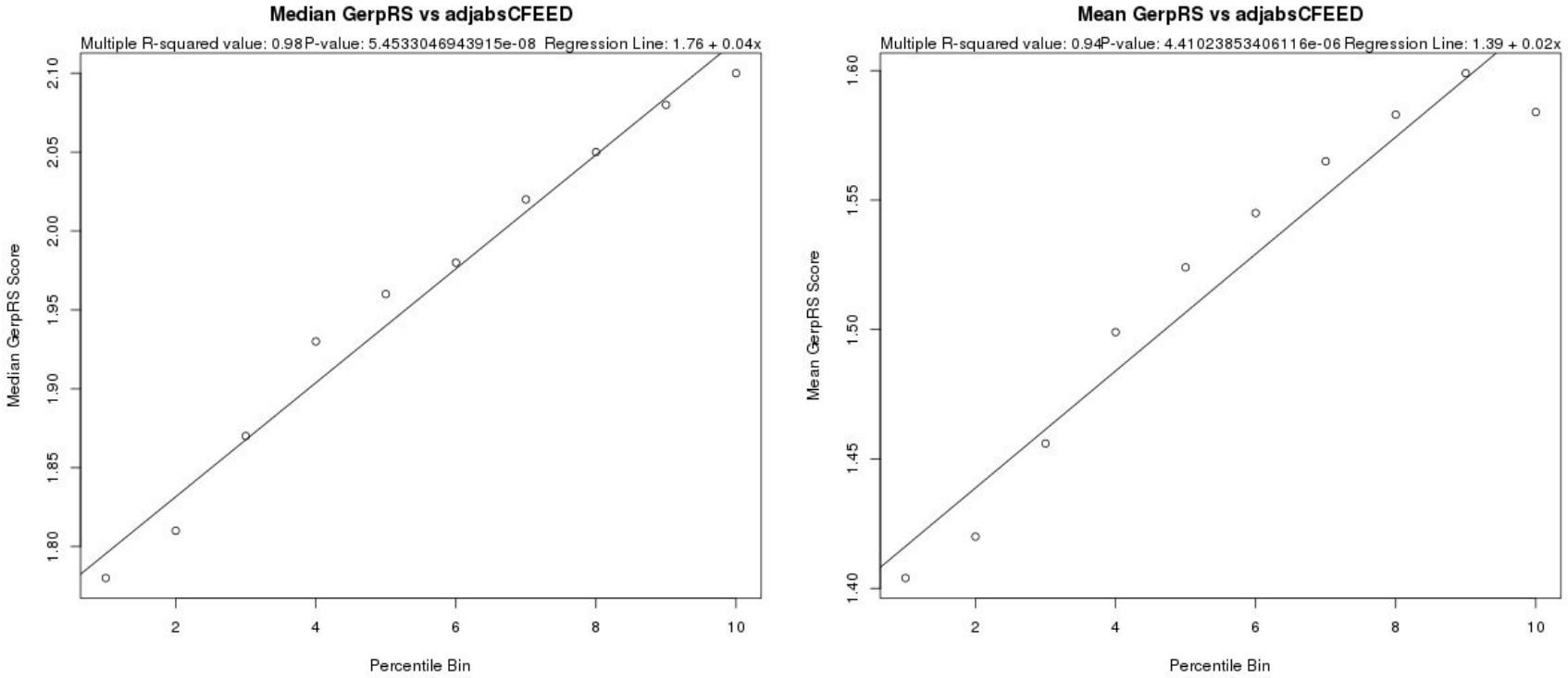
Supplementary Figure 16: % Non-zero Allele Frequency vs. Edit Distance Between Ensembles (EFEED) for All Transcript Regions



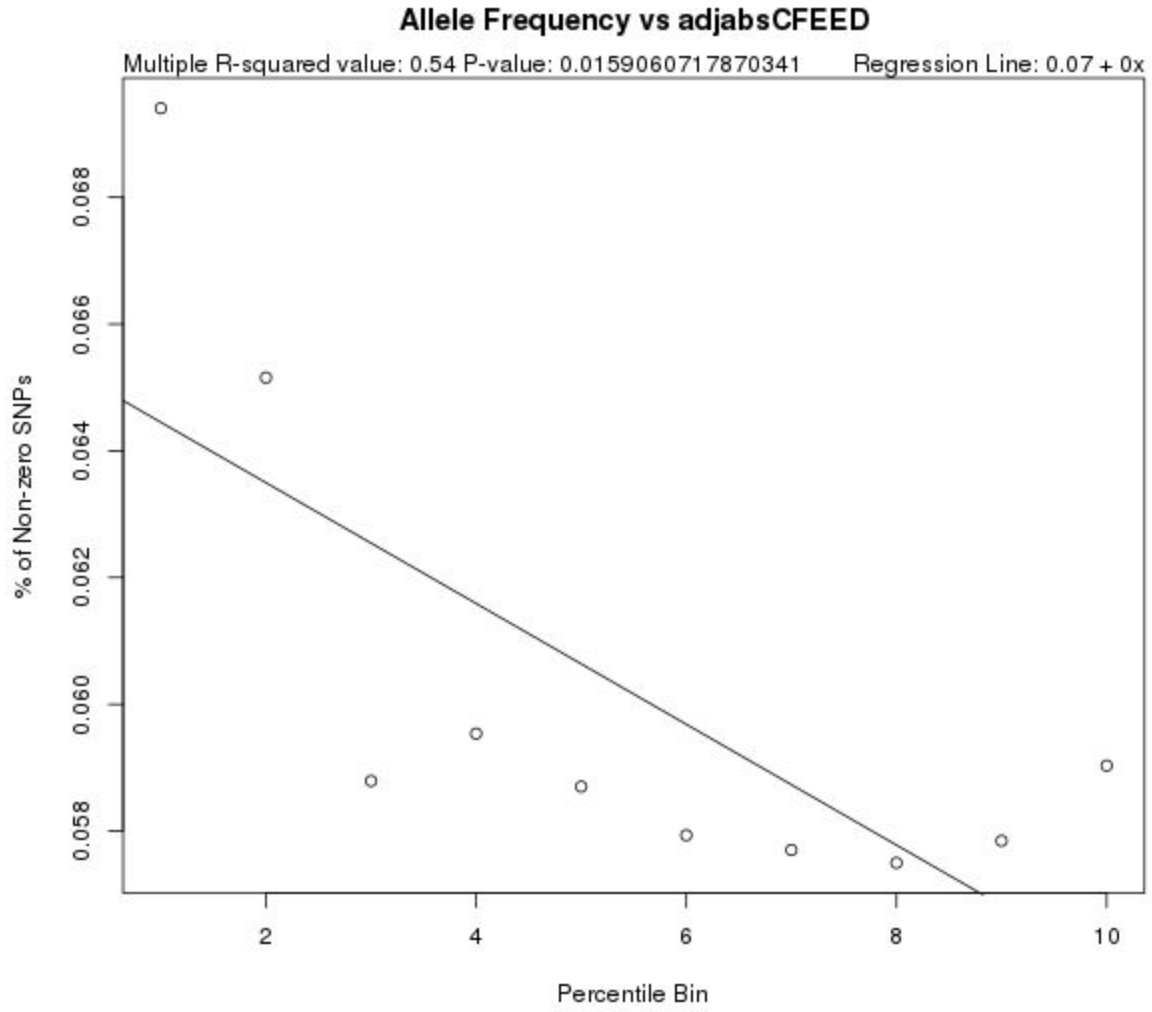
Supplementary Figure 17: Mean/Median GERP Score vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for All Transcript Regions



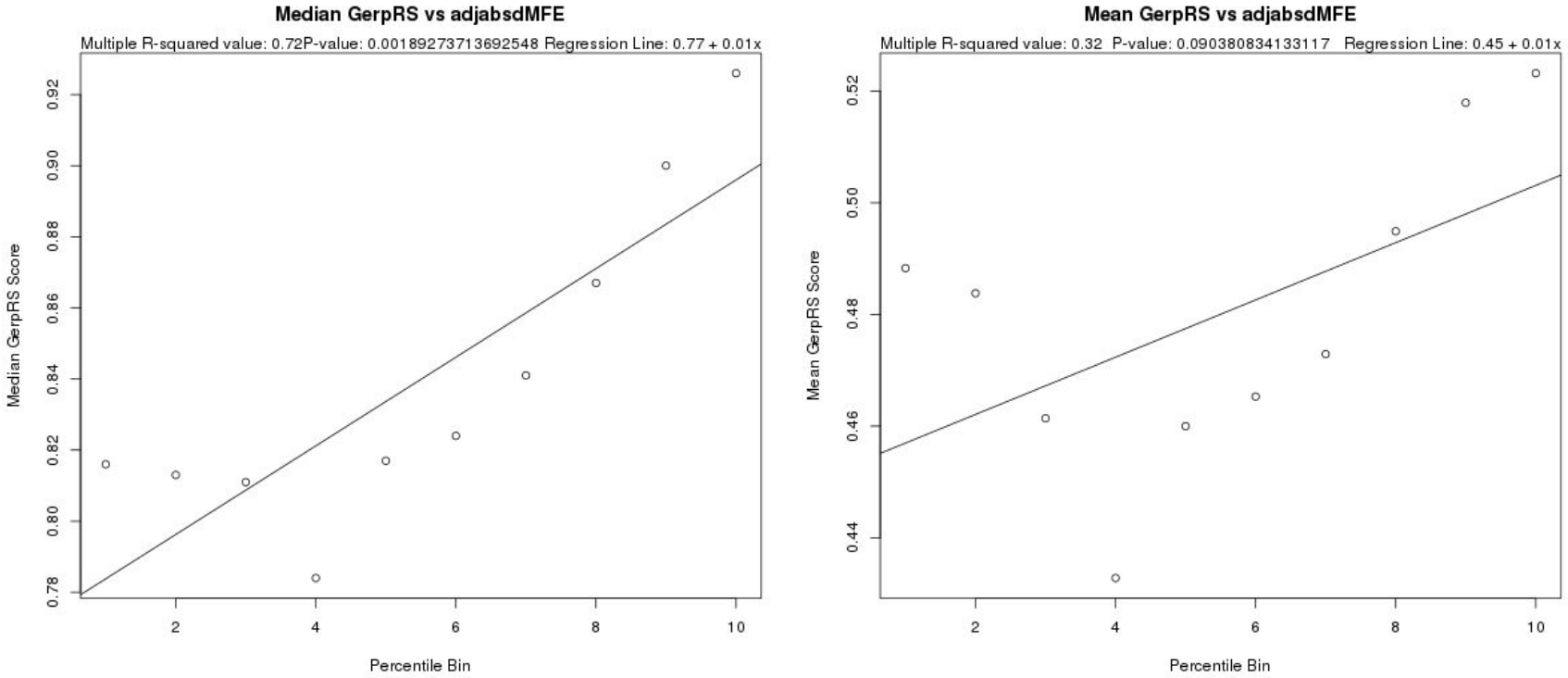
Supplementary Figure 18: % Non-zero Allele Frequency vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for All Transcript Regions



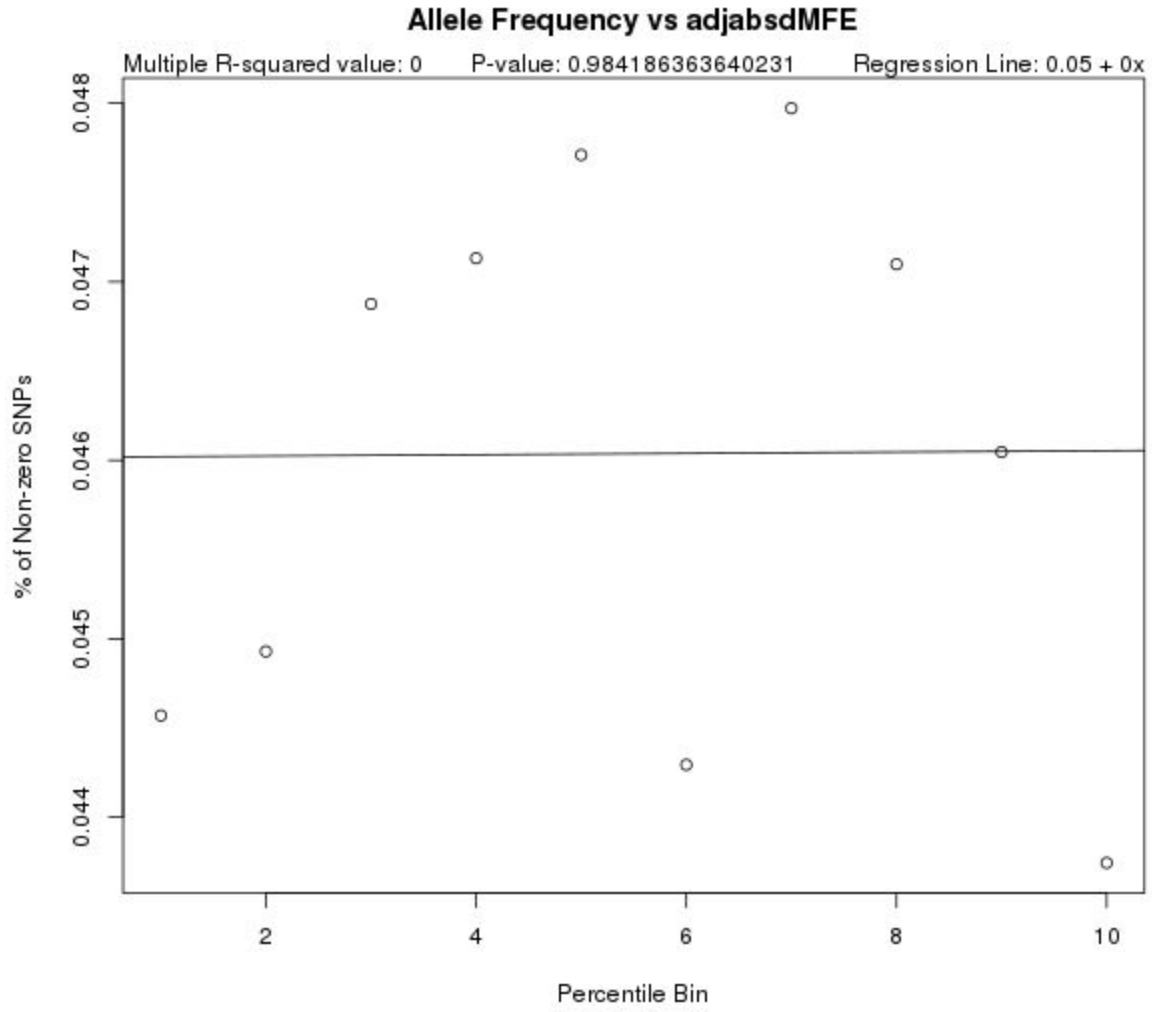
Supplementary Figure 19: Mean/Median GERP Score vs. Edit Distance Between Centroid Structures (CFEED) for All Transcript Regions



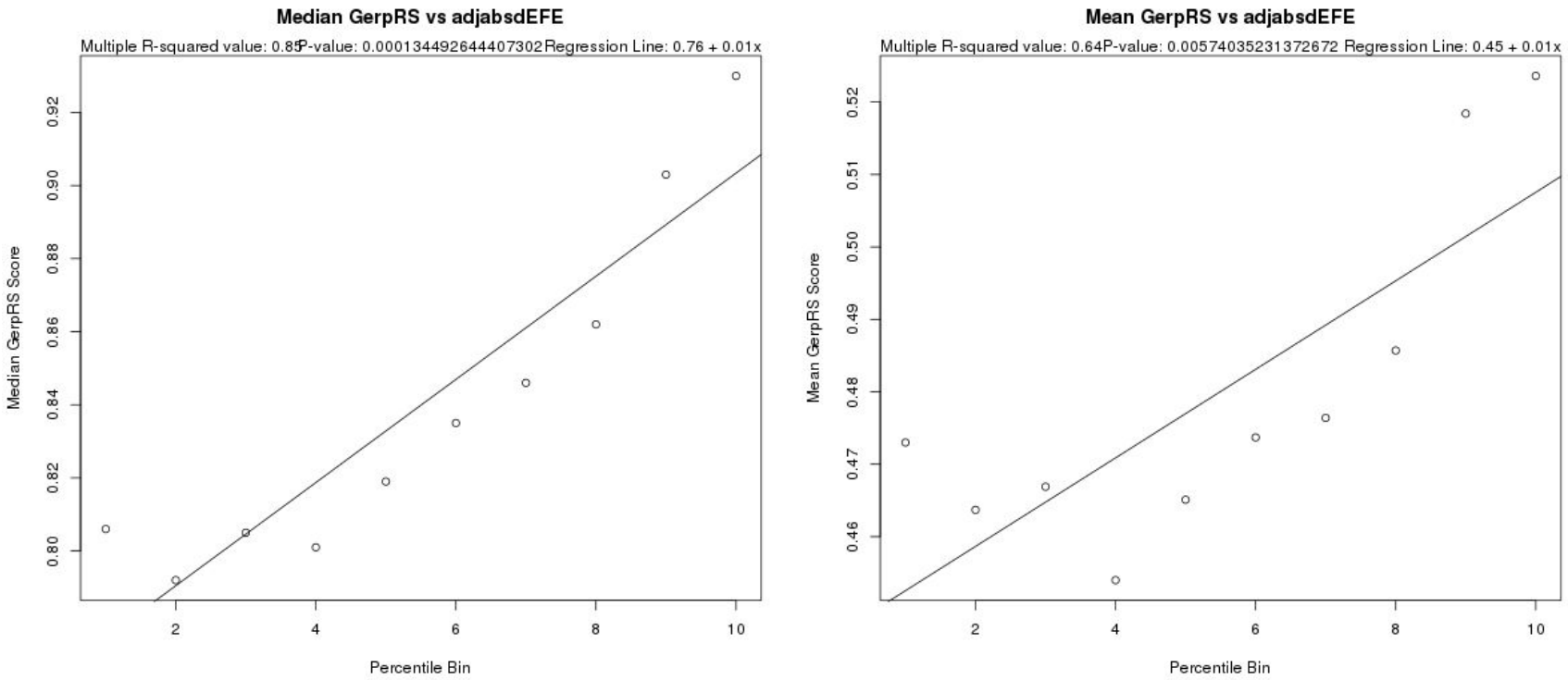
Supplementary Figure 20: % Non-zero Allele Frequency vs. Edit Distance Between Centroid Structures (CFEED) for All Transcript Regions



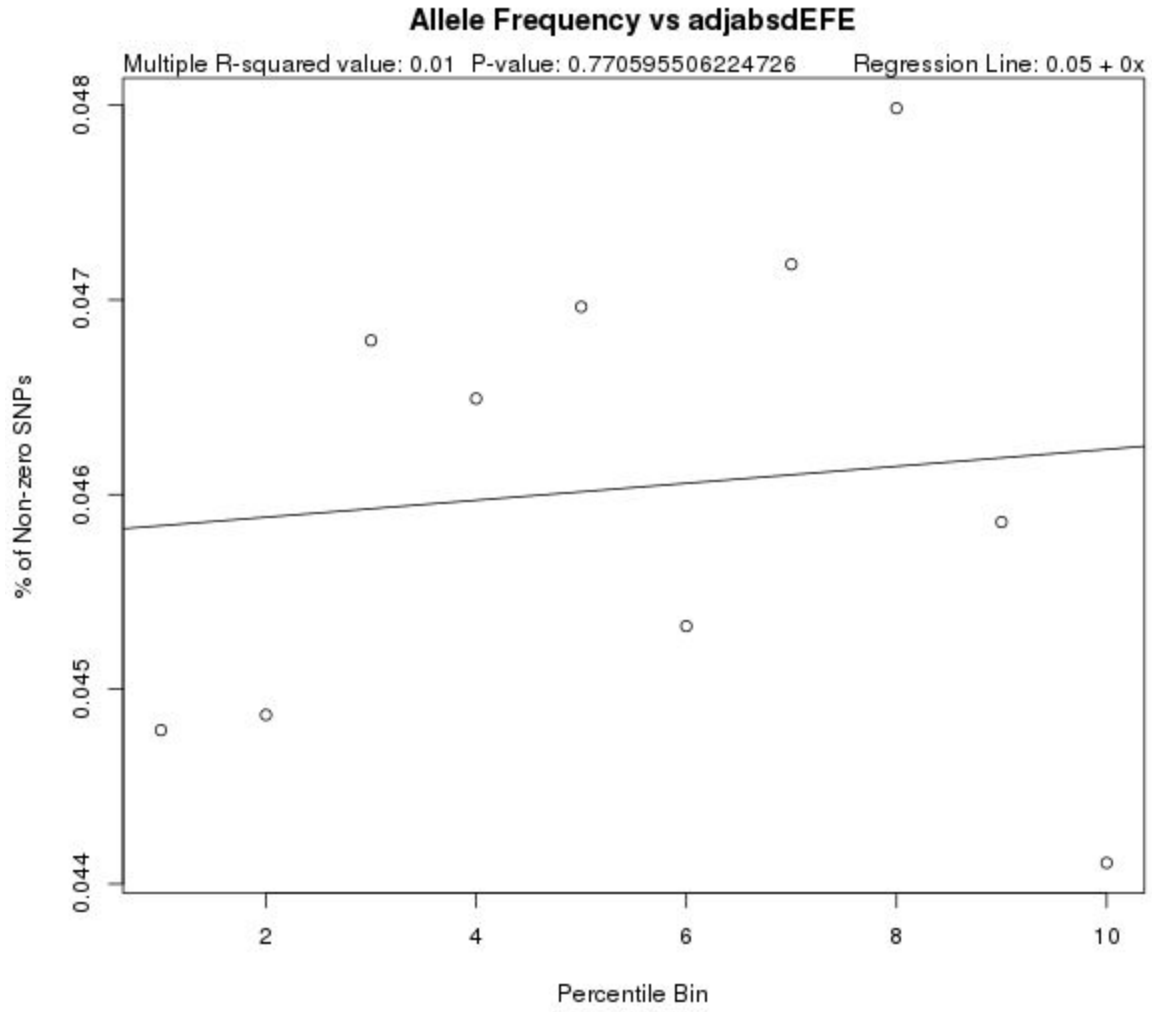
Supplementary Figure 21: Mean/Median GERP Score vs. Binned Change in Minimum Free Energy (dMFE) for 5' UTR Variants



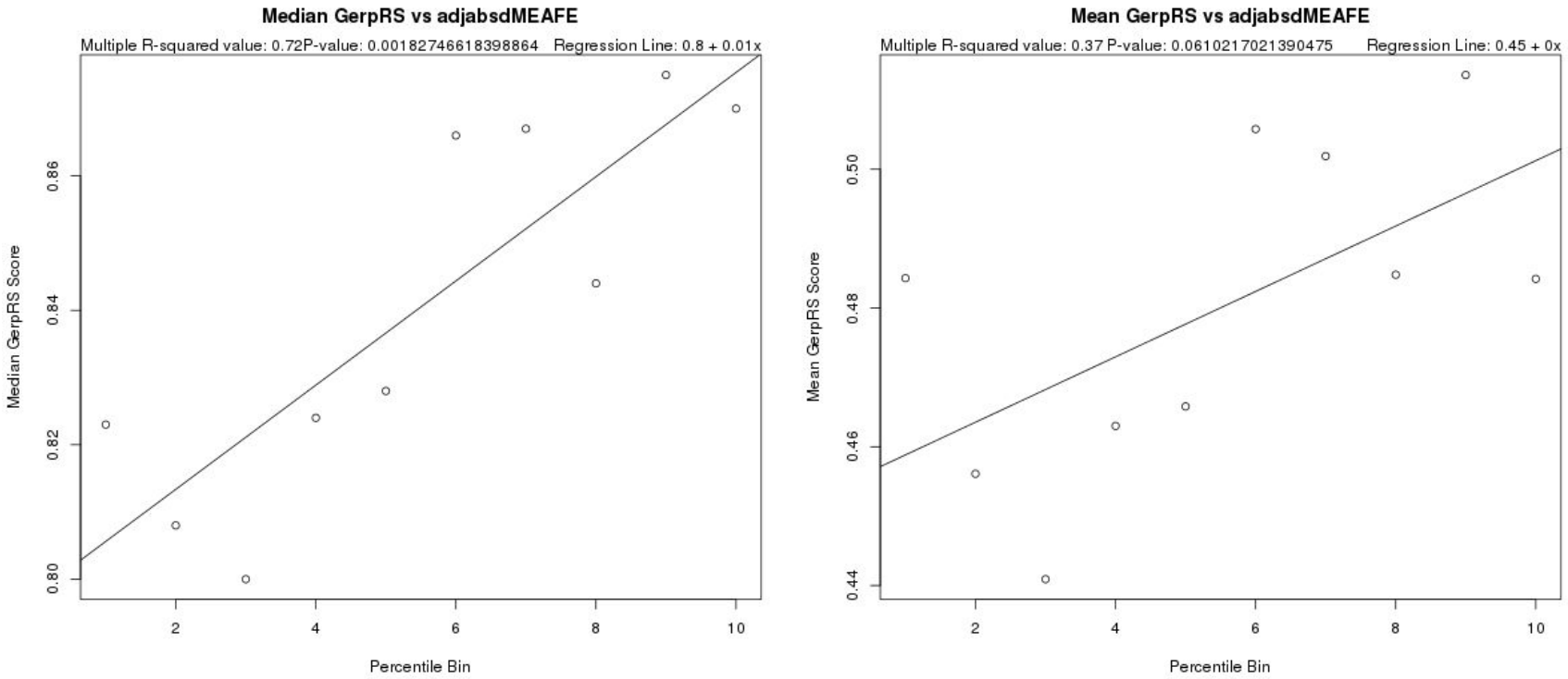
Supplementary Figure 22: % Non-zero Allele Frequency vs. Binned Change in Minimum Free Energy (dMFE) for 5' UTR Variants



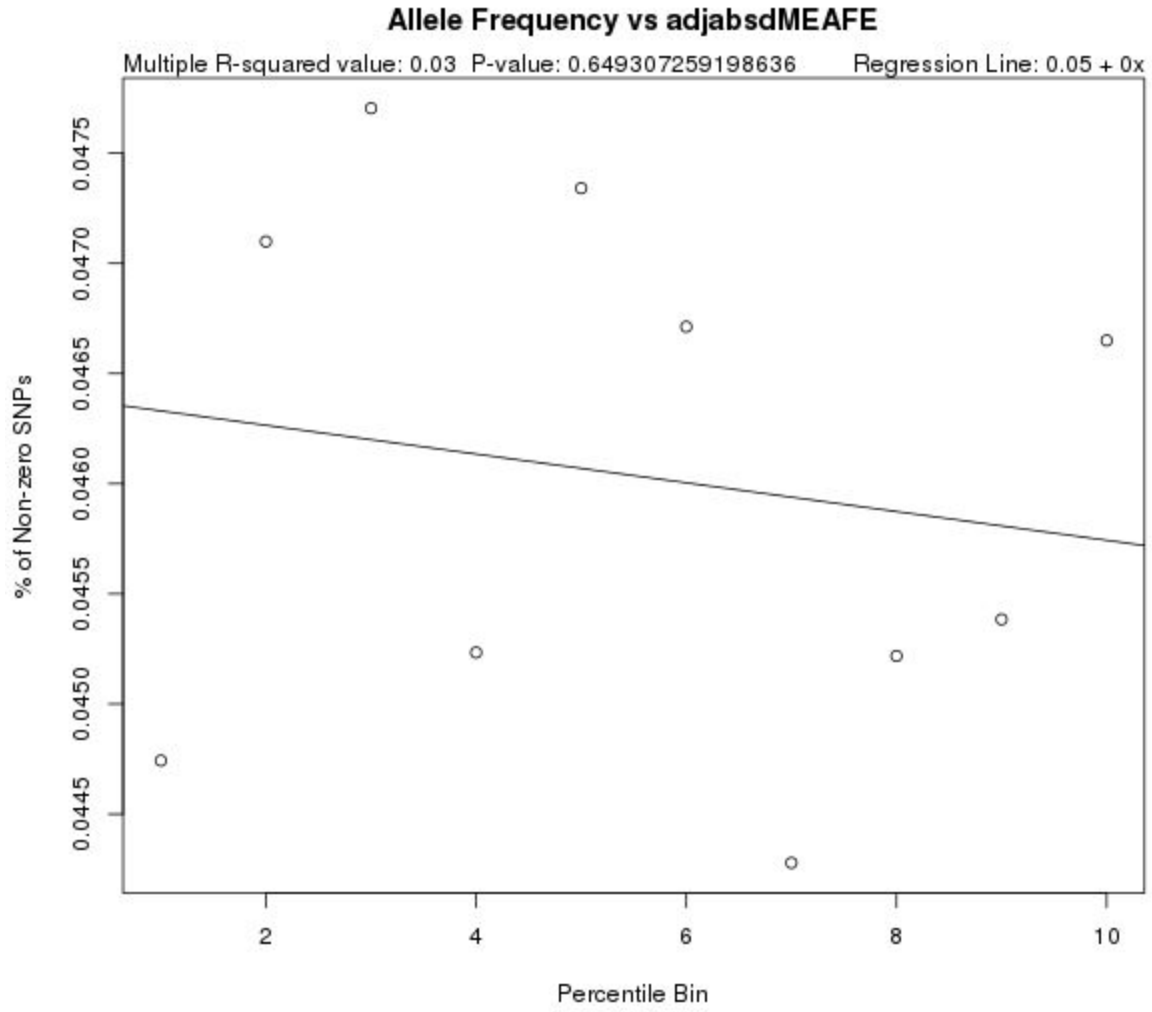
Supplementary Figure 23: Mean/Median GERP Score vs. Binned Change in Ensemble Free Energy (dEFE) for 5' UTR Variants



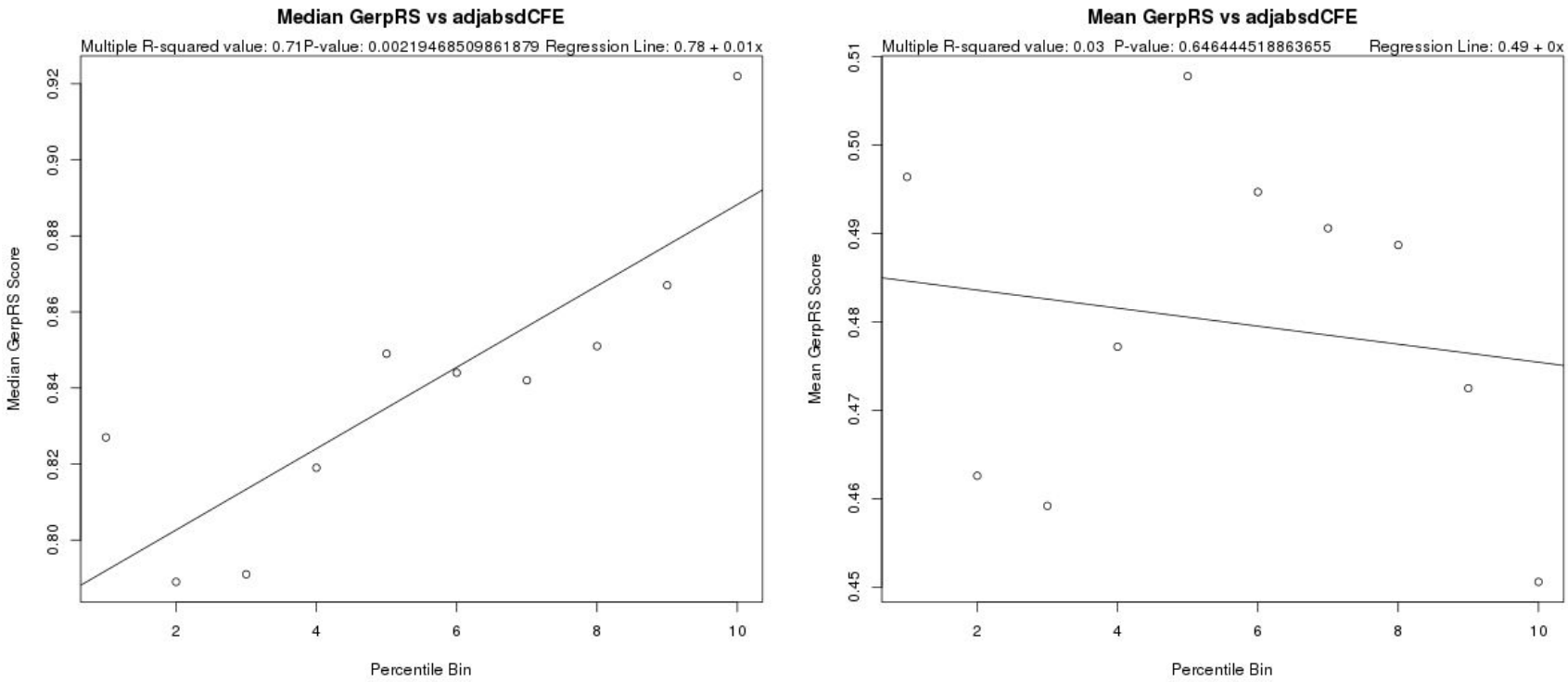
Supplementary Figure 24: % Non-zero Allele Frequency vs. Binned Change in Ensemble Free Energy (dEFE) for 5' UTR Variants



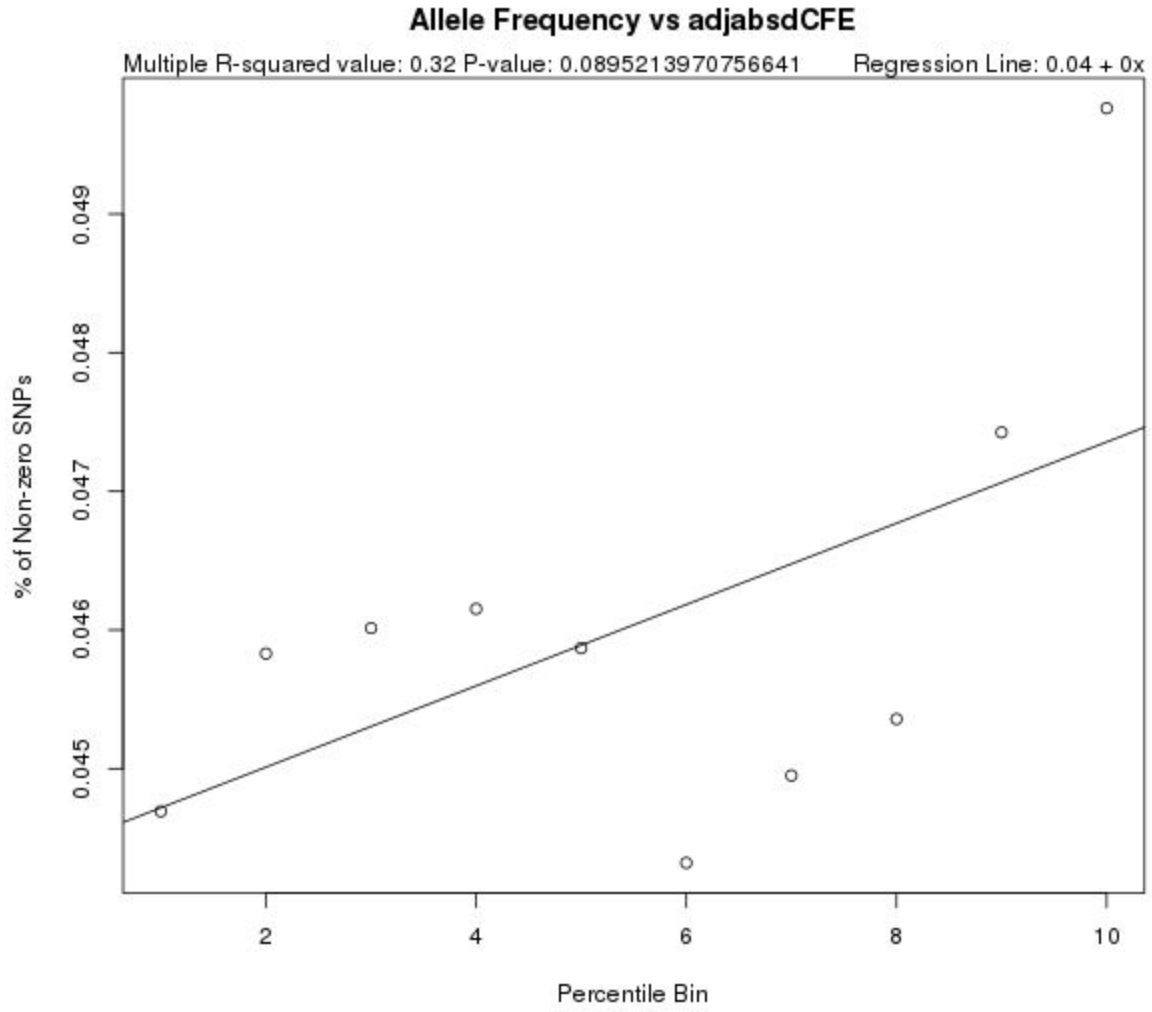
Supplementary Figure 25: Mean/Median GERP Score vs. Binned Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for 5' UTR Variants



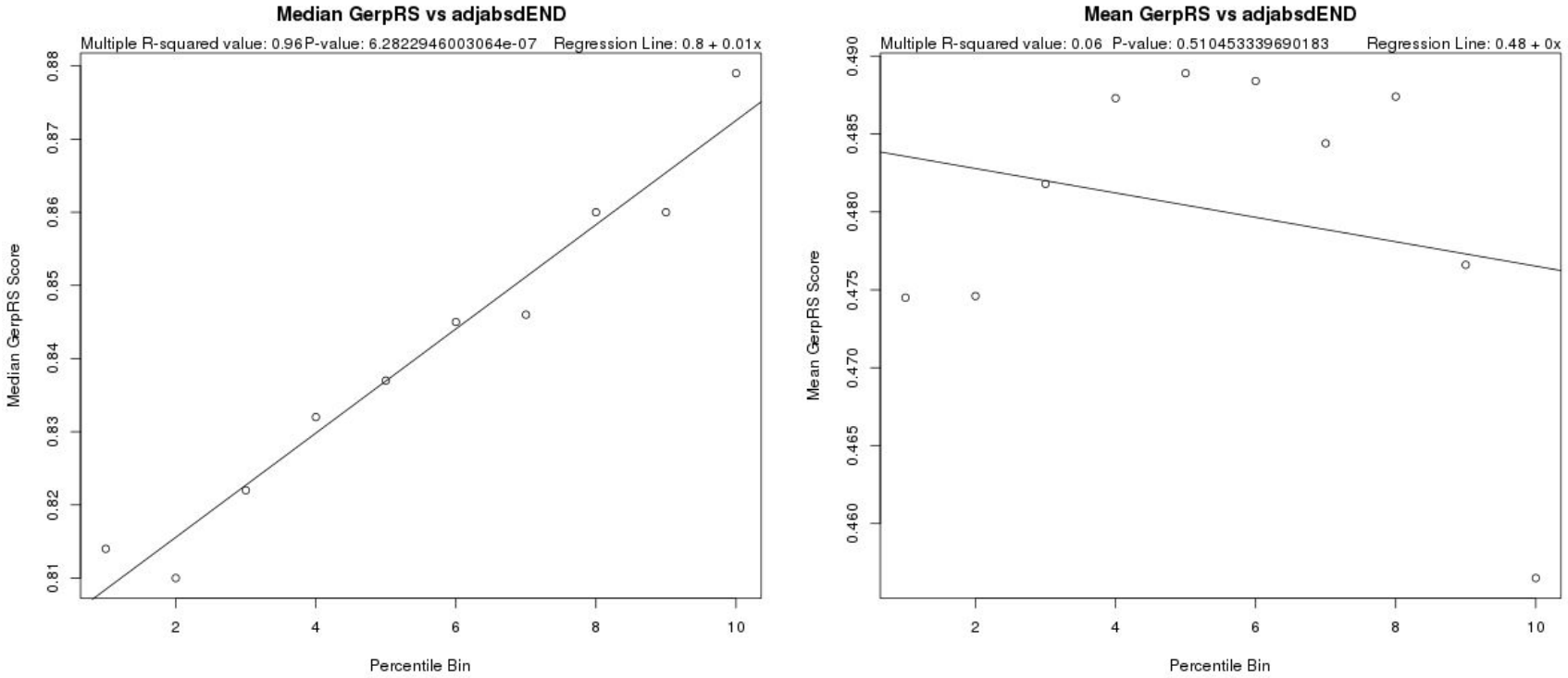
Supplementary Figure 26: % Non-zero Allele Frequency vs. Binned Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for 5' UTR Variants



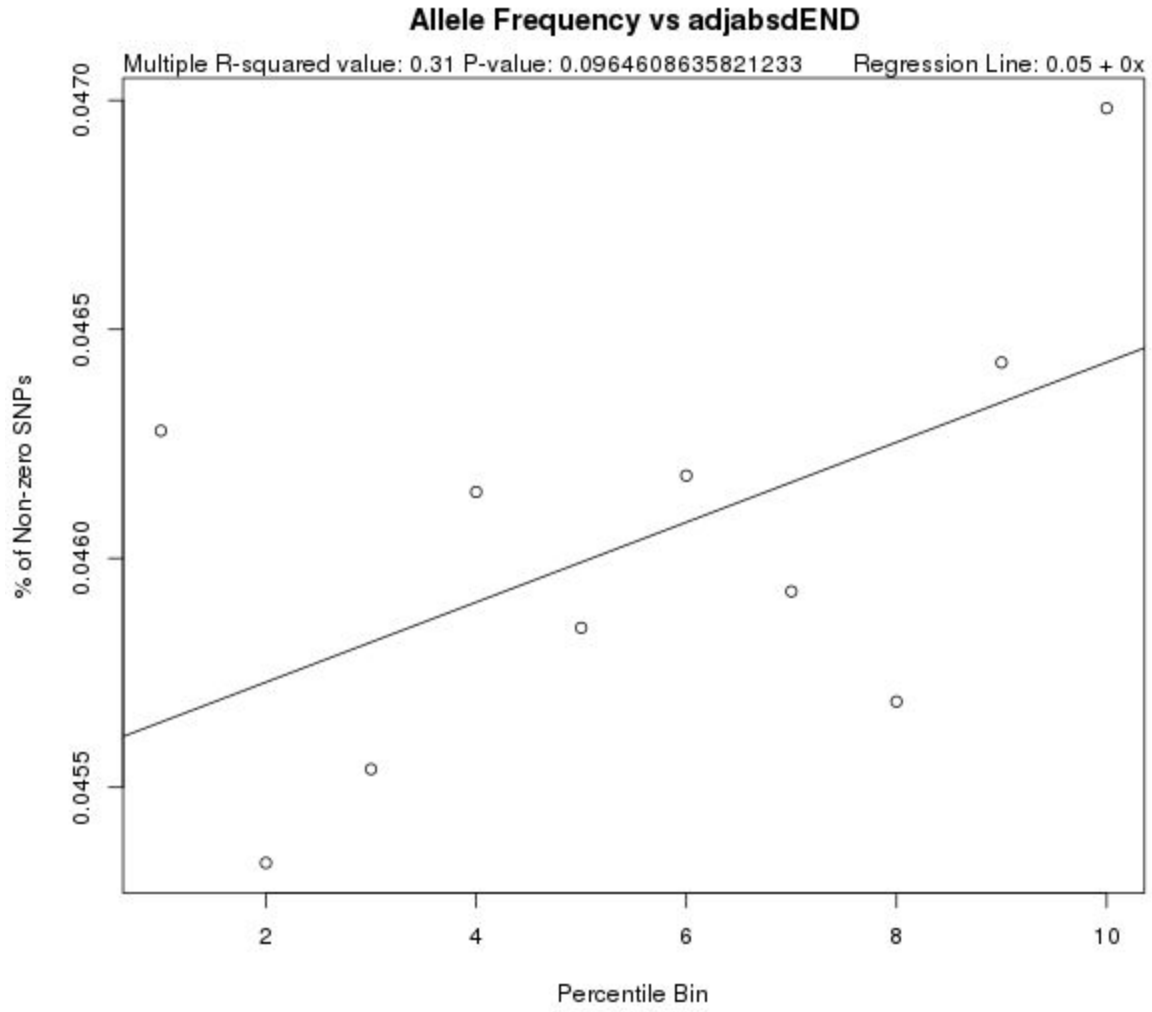
Supplementary Figure 27: Mean/Median GERP Score vs. Binned Change in Free Energy of the Centroid (dCFE) for 5' UTR Variants



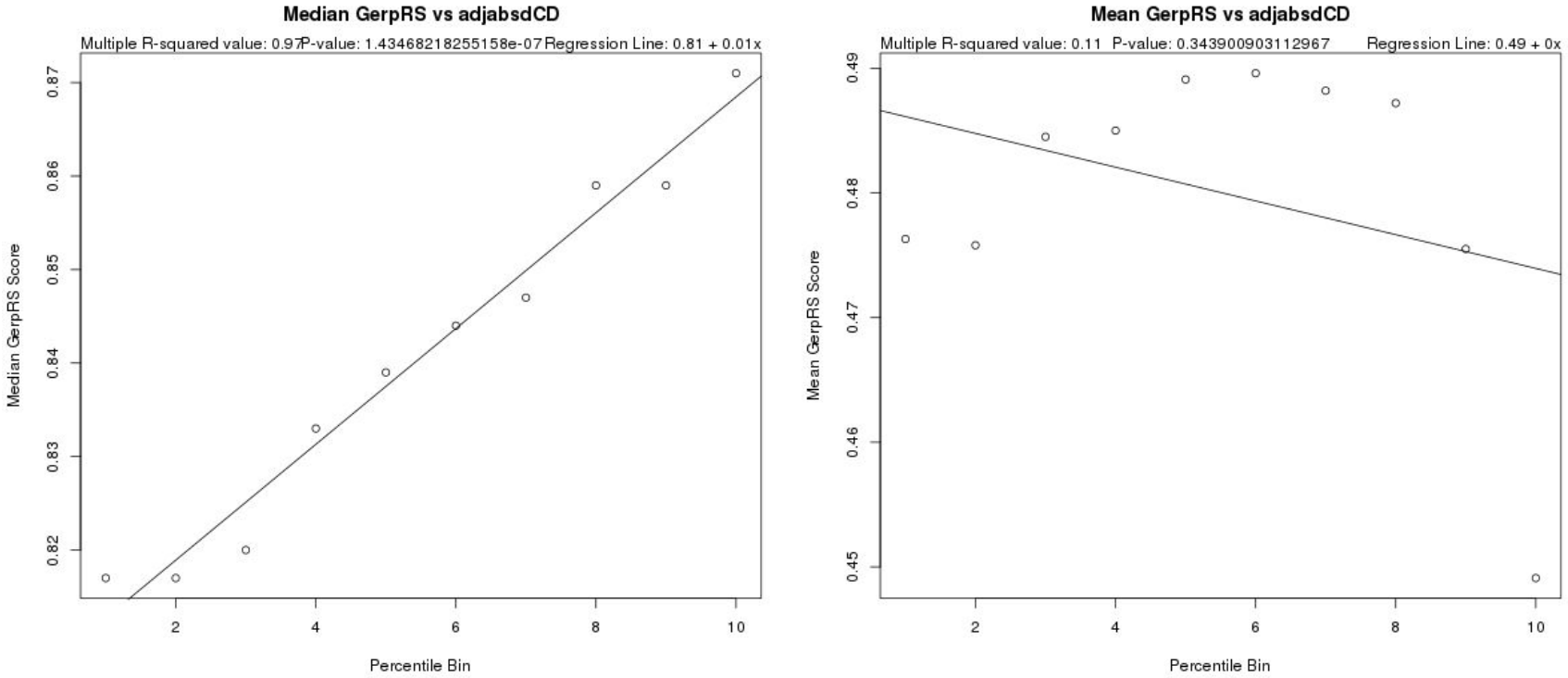
Supplementary Figure 28: % Non-zero Allele Frequency vs. Binned Change in Free Energy of the Centroid (dCFE) for 5' UTR Variants



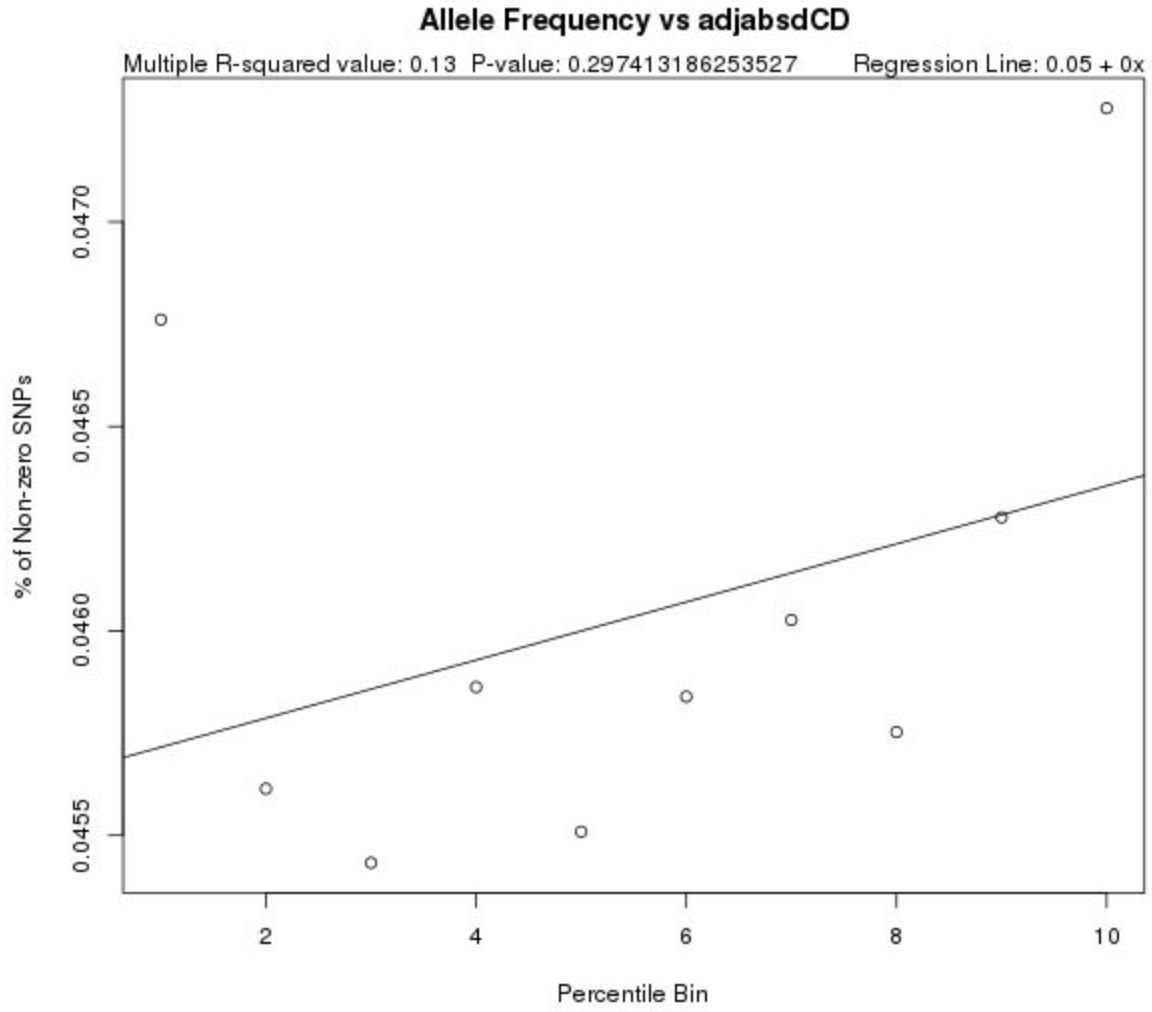
Supplementary Figure 29: Mean/Median GERP Score vs. Binned Change in Ensemble Diversity (dEND) for 5' UTR Variants



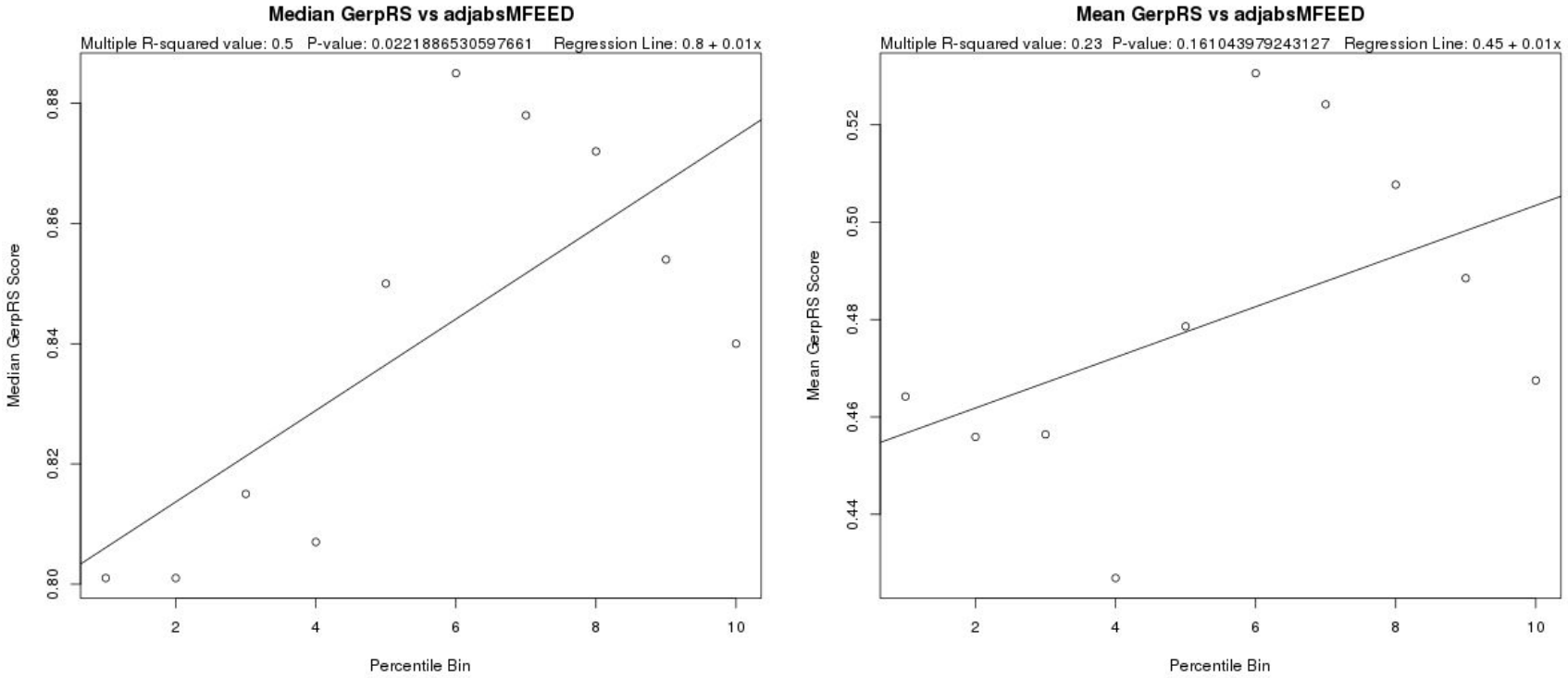
Supplementary Figure 30: % Non-zero Allele Frequency vs. Binned Change in Ensemble Diversity (dEND) for 5' UTR Variants



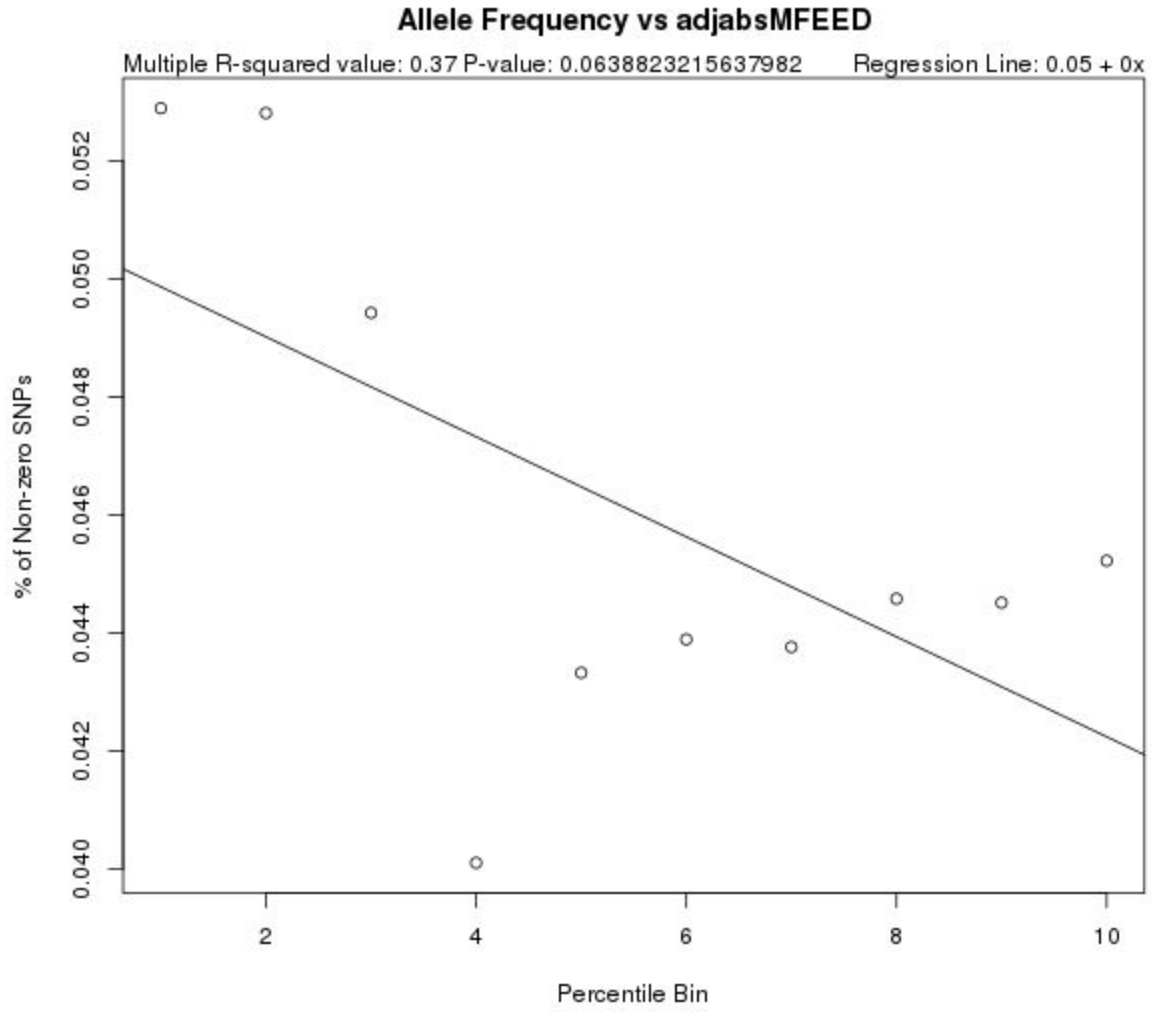
Supplementary Figure 31: Mean/Median GERP Score vs. Binned Change in Distance of the Ensemble of Structures to the Centroid (dCD) for 5' UTR Variants



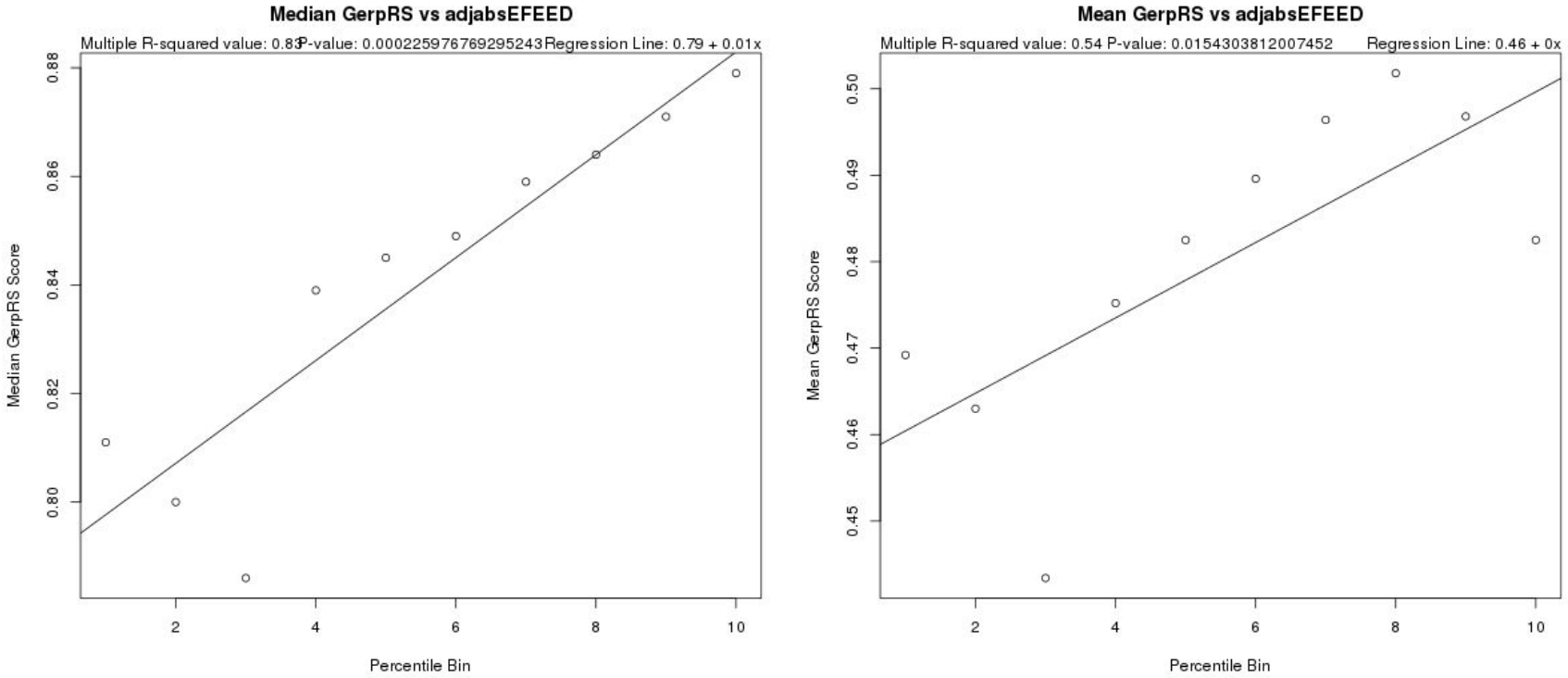
Supplementary Figure 32: % Non-zero Allele Frequency vs. Binned Change in Distance of the Ensemble of Structures to the Centroid (dCD) for 5' UTR Variants



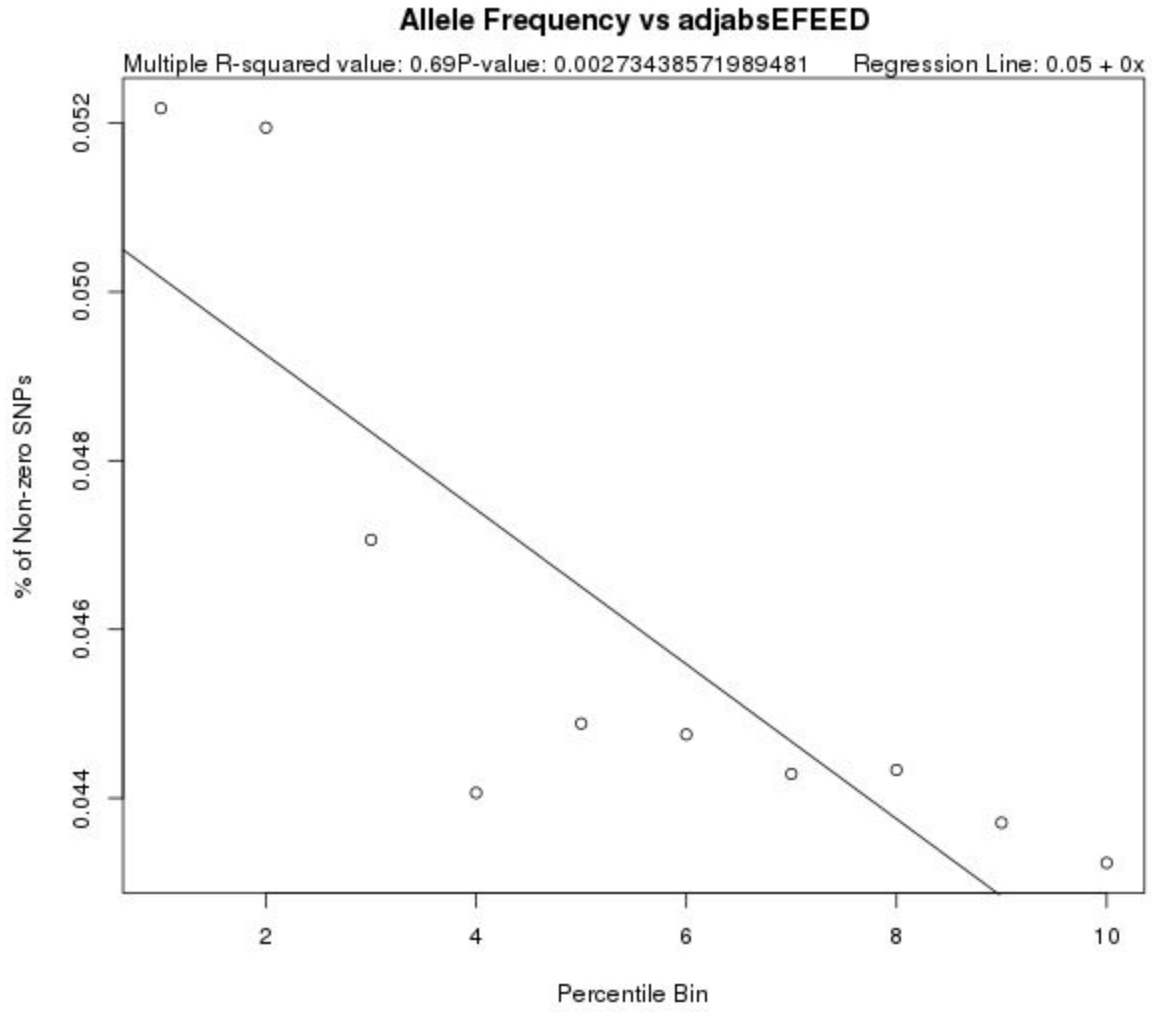
Supplementary Figure 33: Mean/Median GERP Score vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for 5' UTR Variants



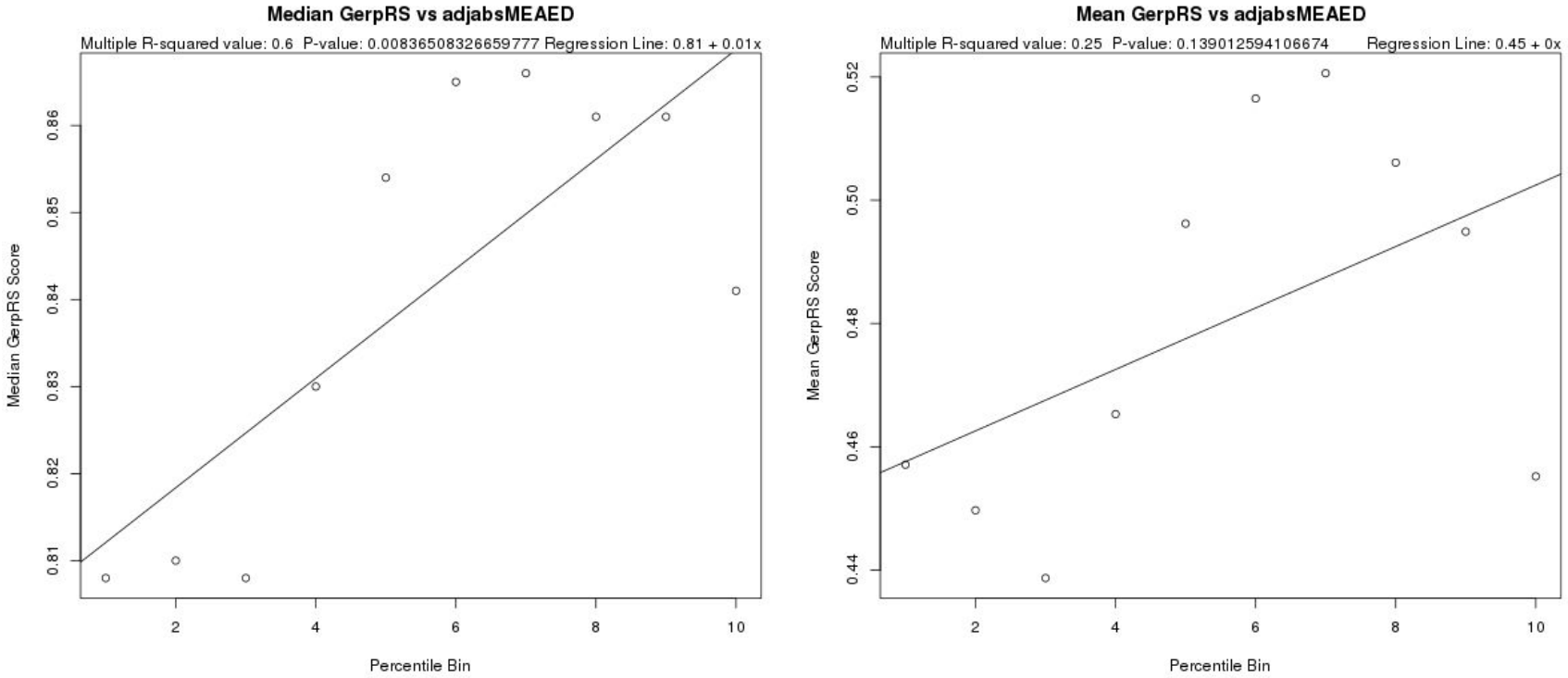
Supplementary Figure 34: % Non-zero Allele Frequency vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for 5' UTR Variants



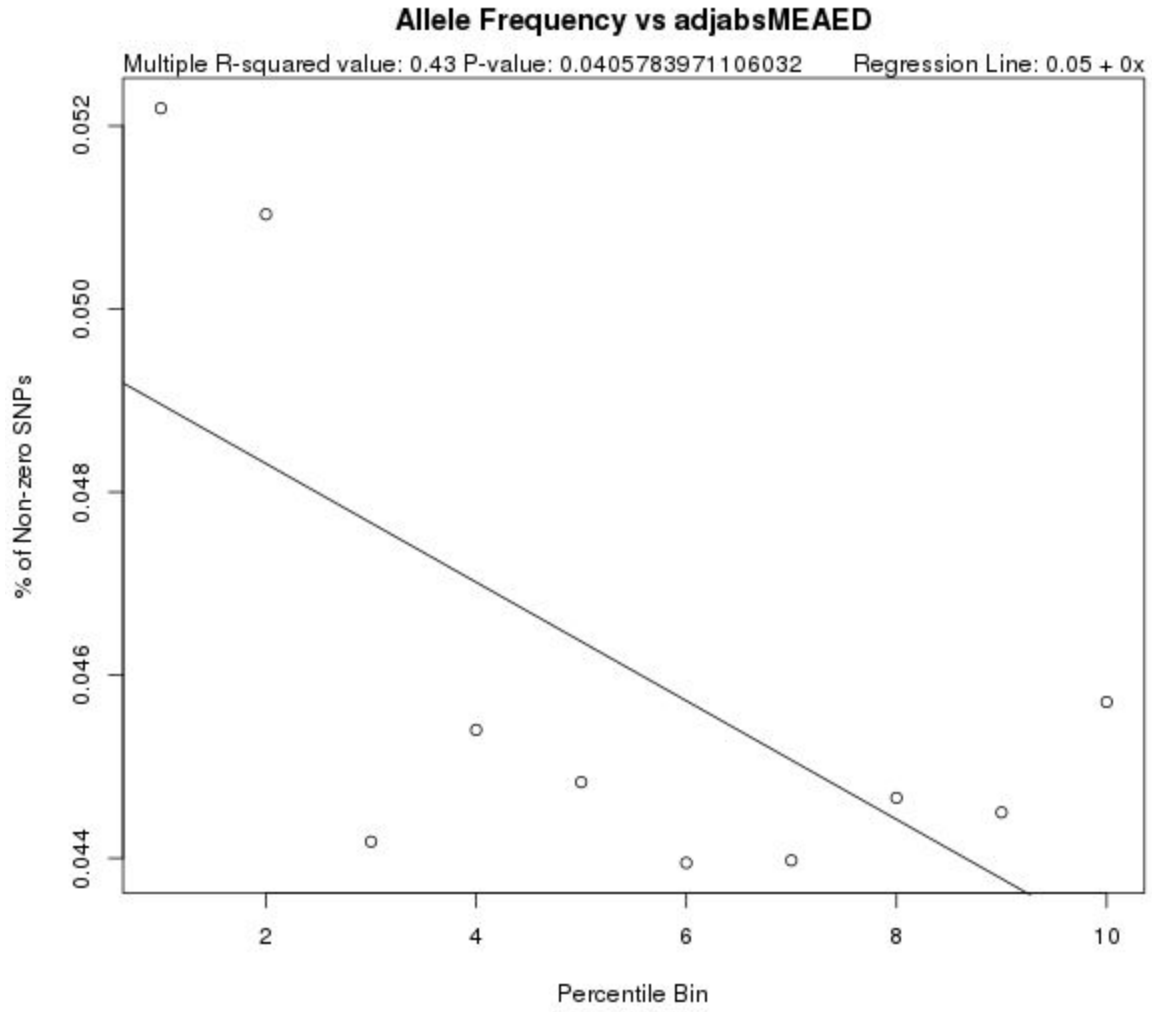
Supplementary Figure 35: Mean/Median GERP Score vs. Edit Distance Between Ensembles (EFEED) for 5' UTR Variants



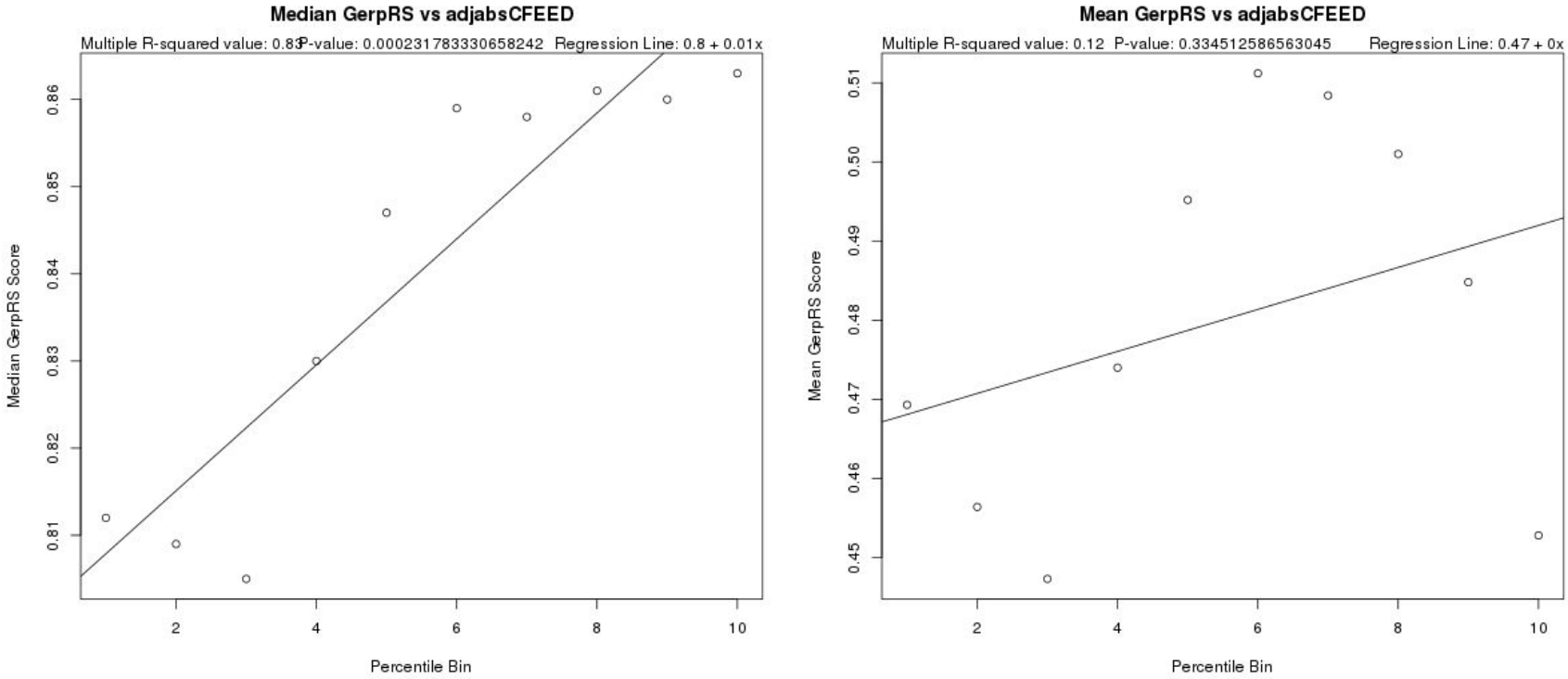
Supplementary Figure 36: % Non-zero Allele Frequency vs. Edit Distance Between Ensembles (EFEED) for 5' UTR Variants



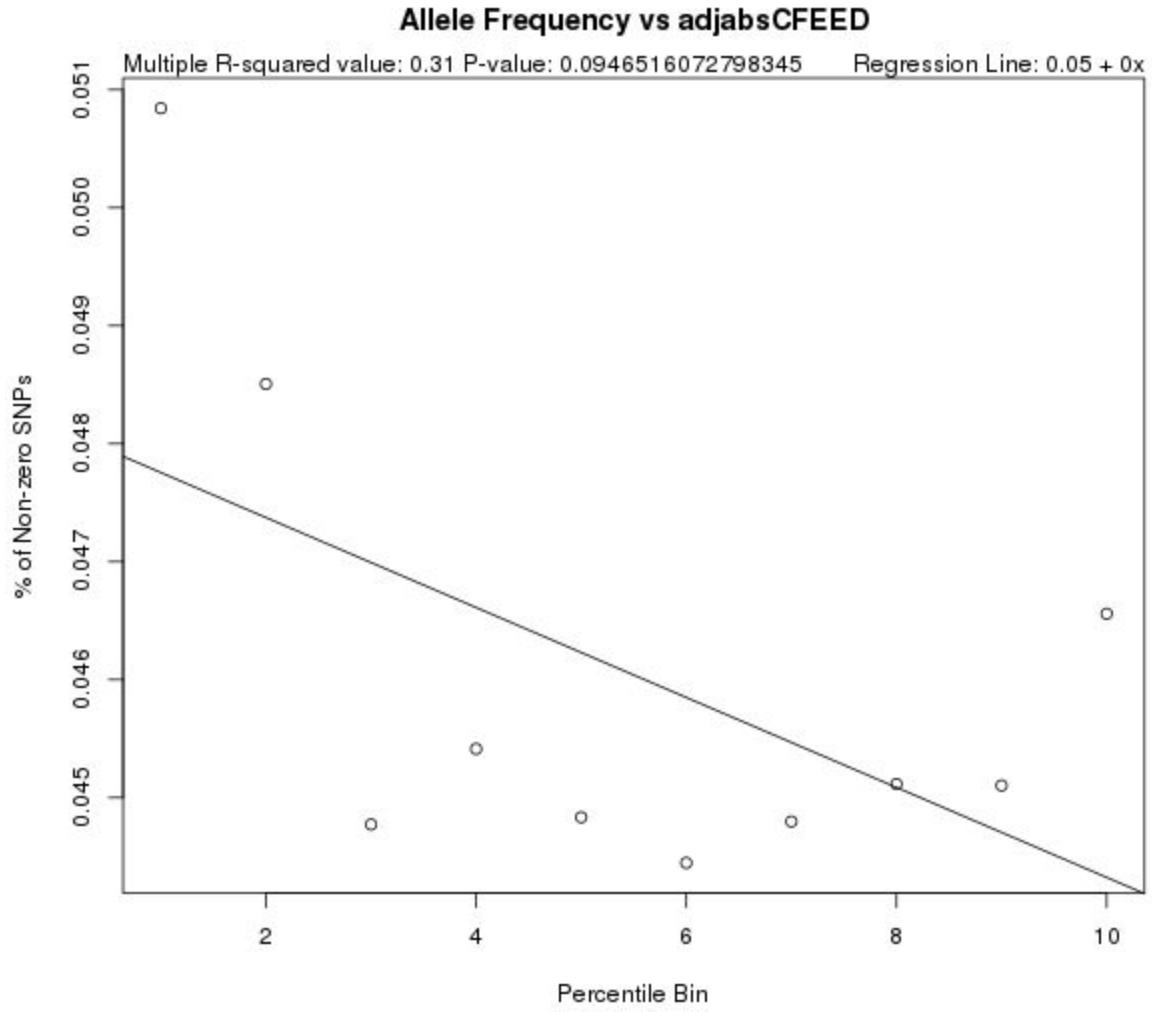
Supplementary Figure 37: Mean/Median GERP Score vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for 5' UTR Variants



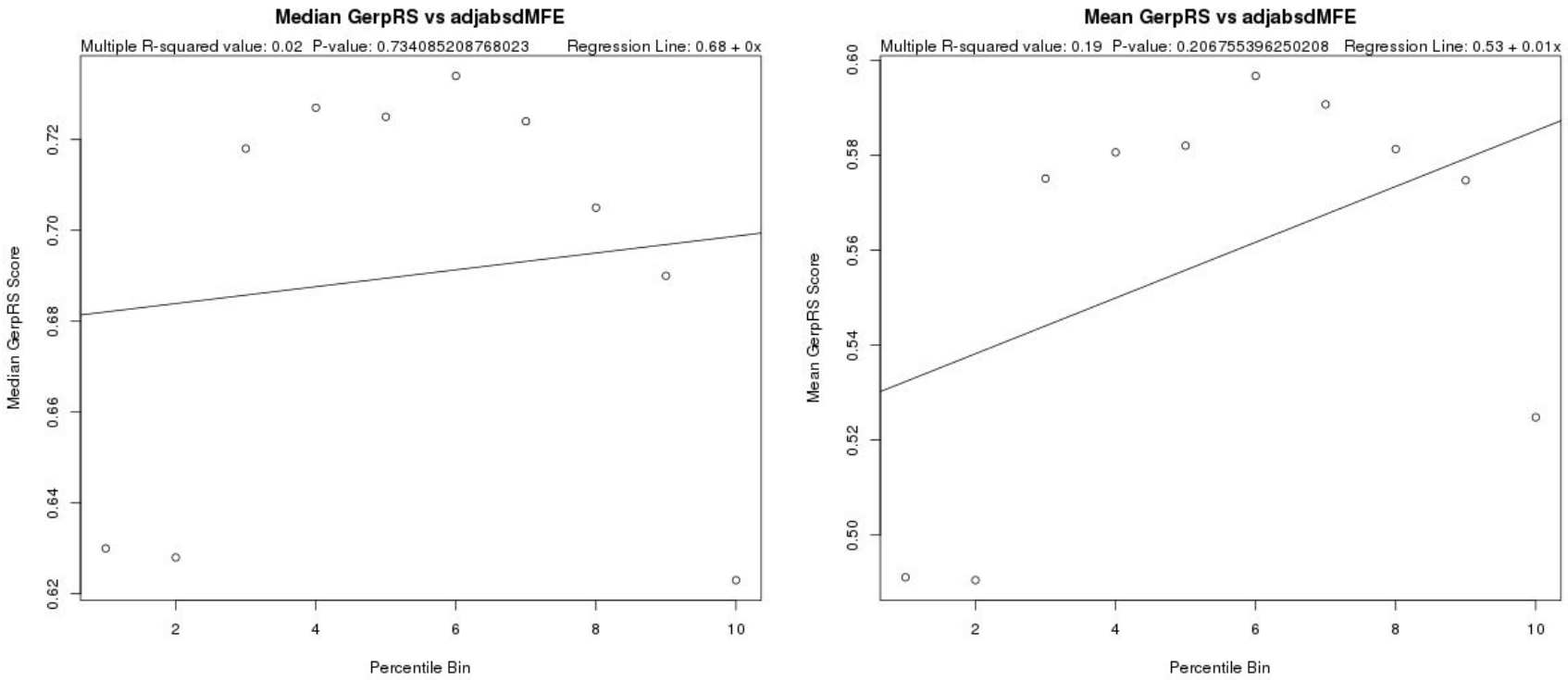
Supplementary Figure 38: % Non-zero Allele Frequency vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for 5' UTR Variants



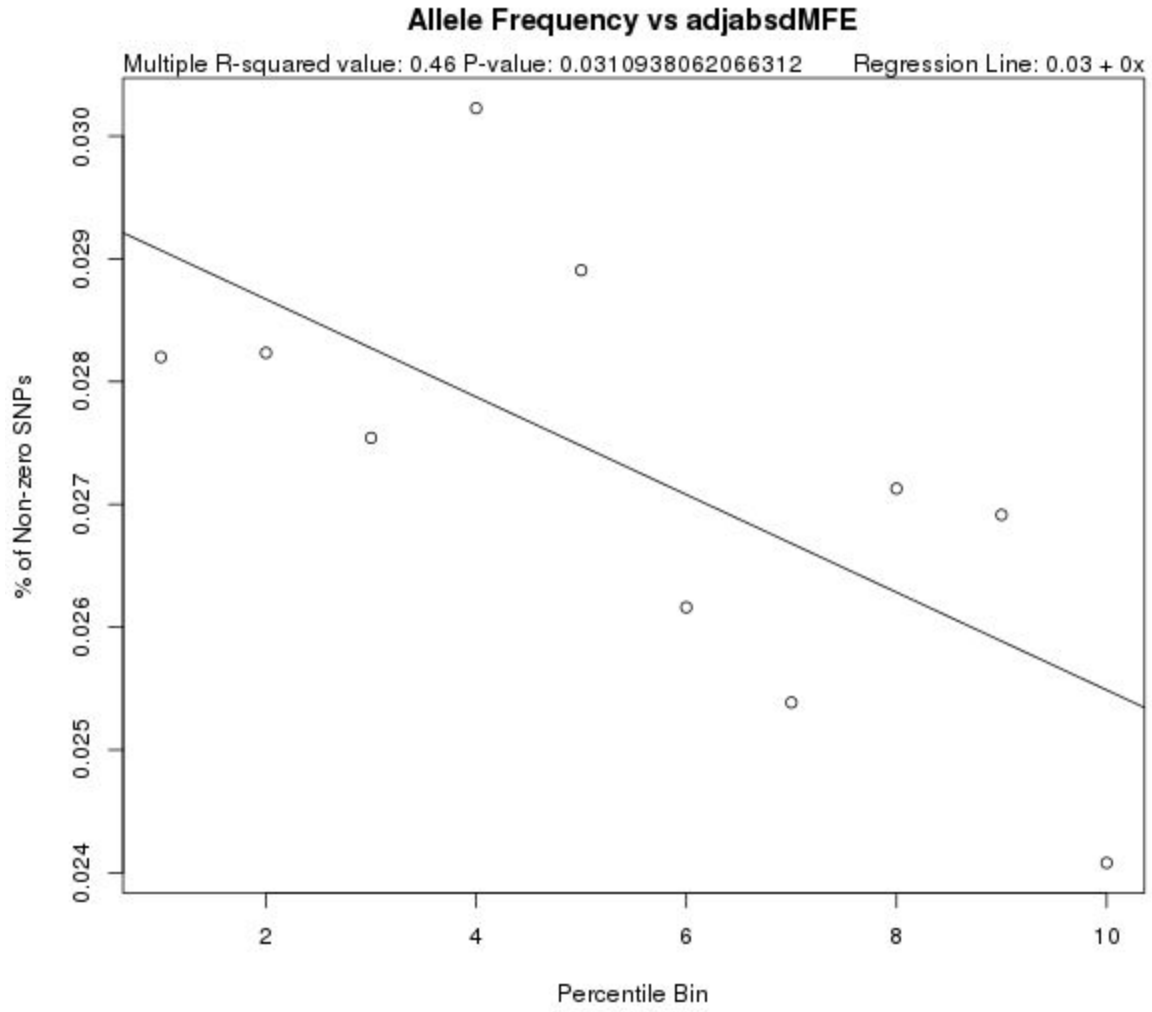
Supplementary Figure 39: Mean/Median GERP Score vs. Edit Distance Between Centroid Structures (CFEED) for 5' UTR Variants



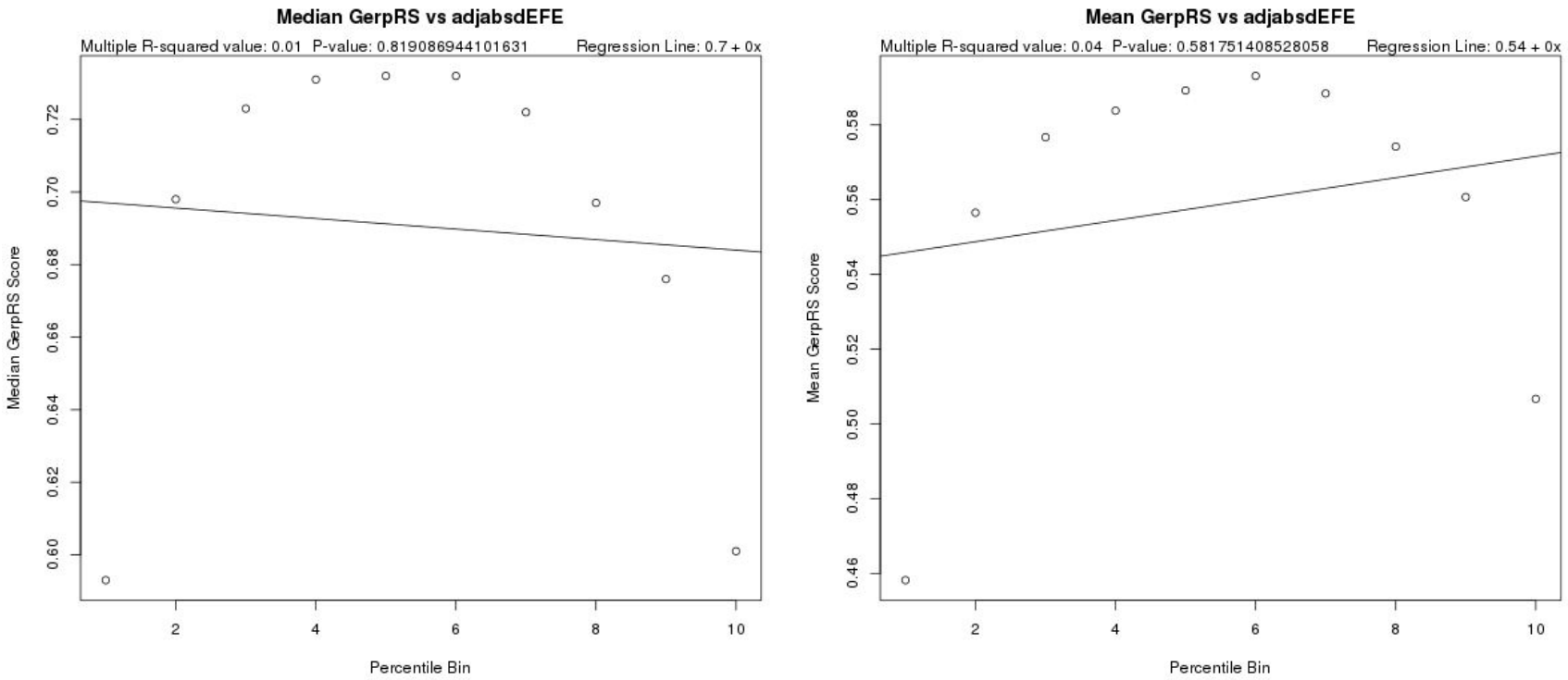
Supplementary Figure 40: % Non-zero Allele Frequency vs. Edit Distance Between Centroid Structures (CFEED) for 5' UTR Variants



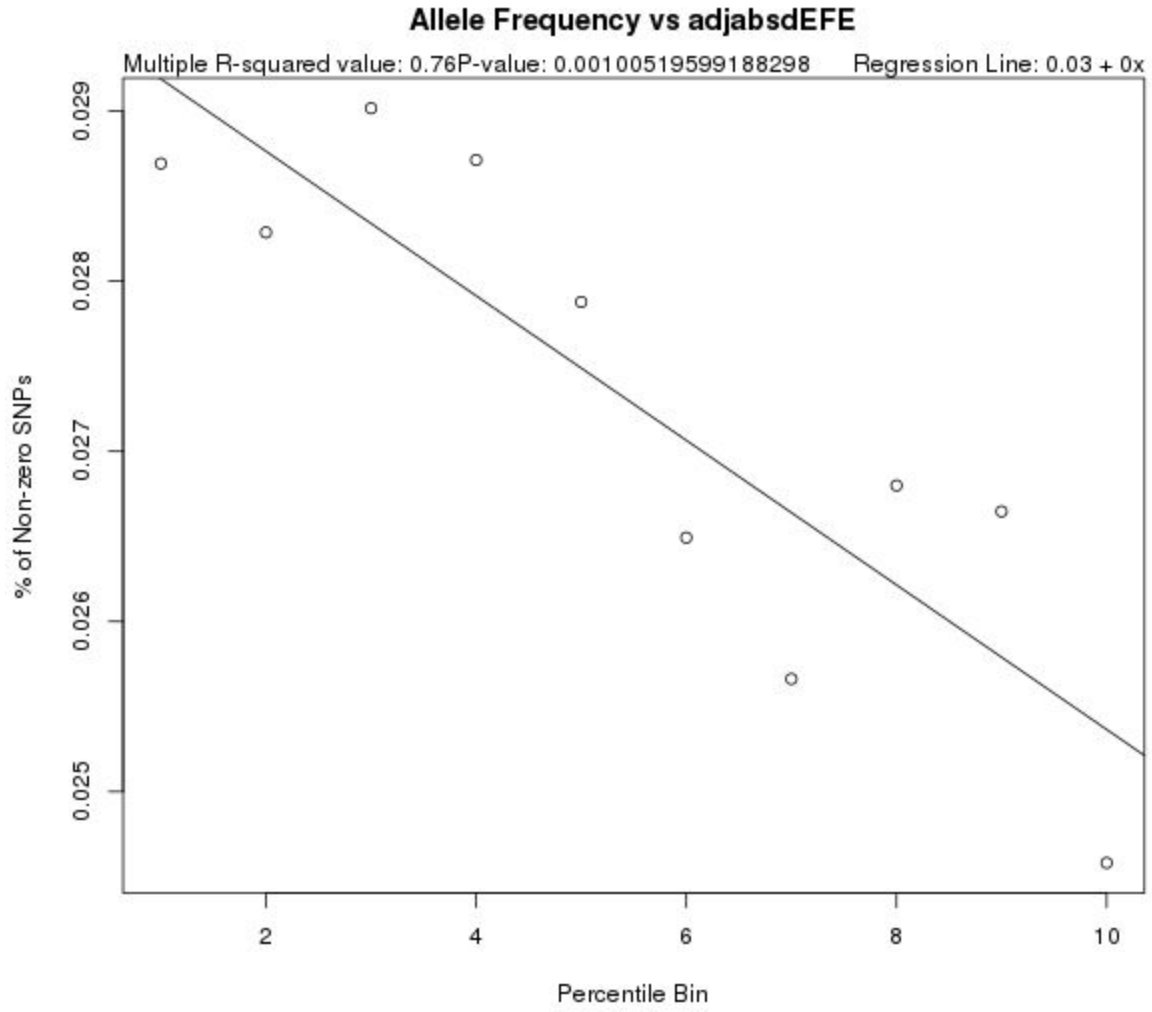
Supplementary Figure 41: Mean/Median GERP Score vs. Binned Change in Minimum Free Energy (dMFE) for 3' UTR Variants



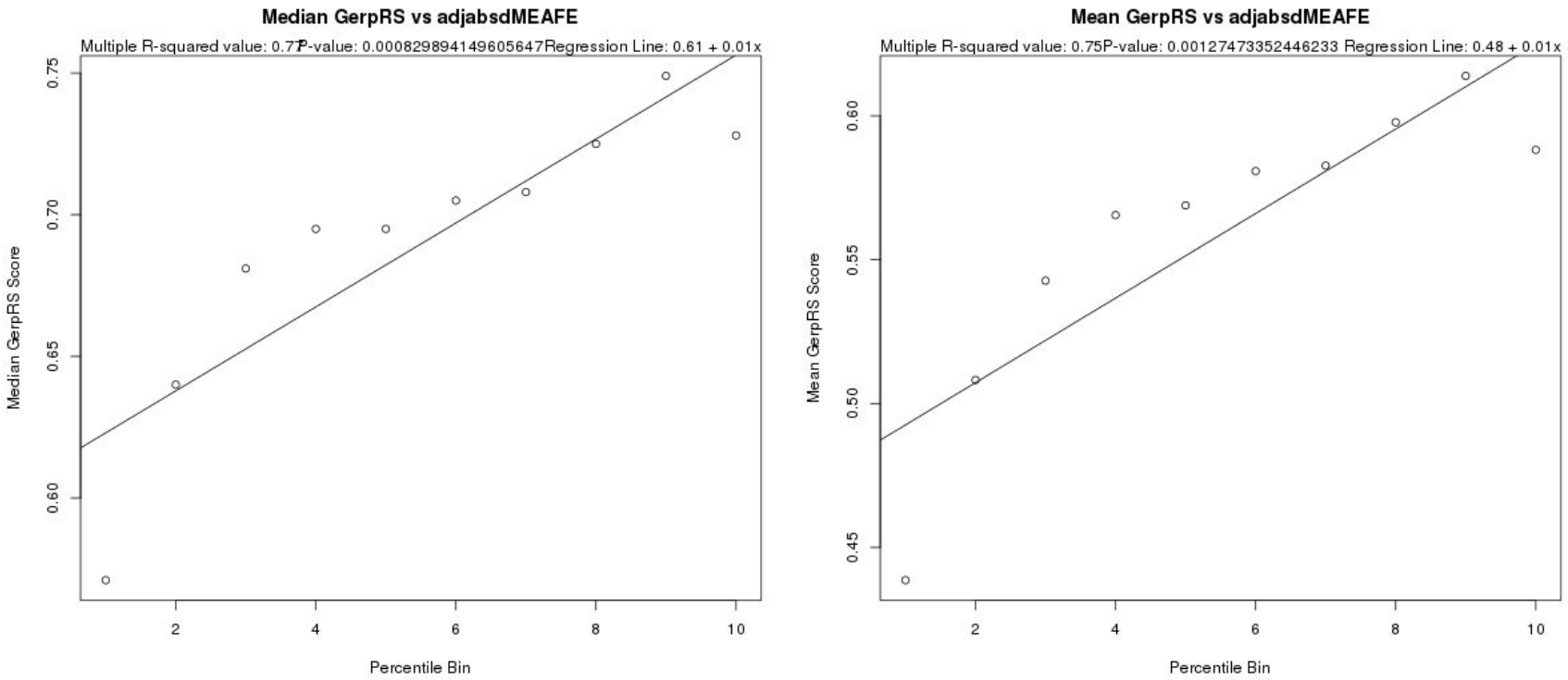
Supplementary Figure 42: % Non-zero Allele Frequency vs. Binned Change in Minimum Free Energy (dMFE) for 3' UTR Variants



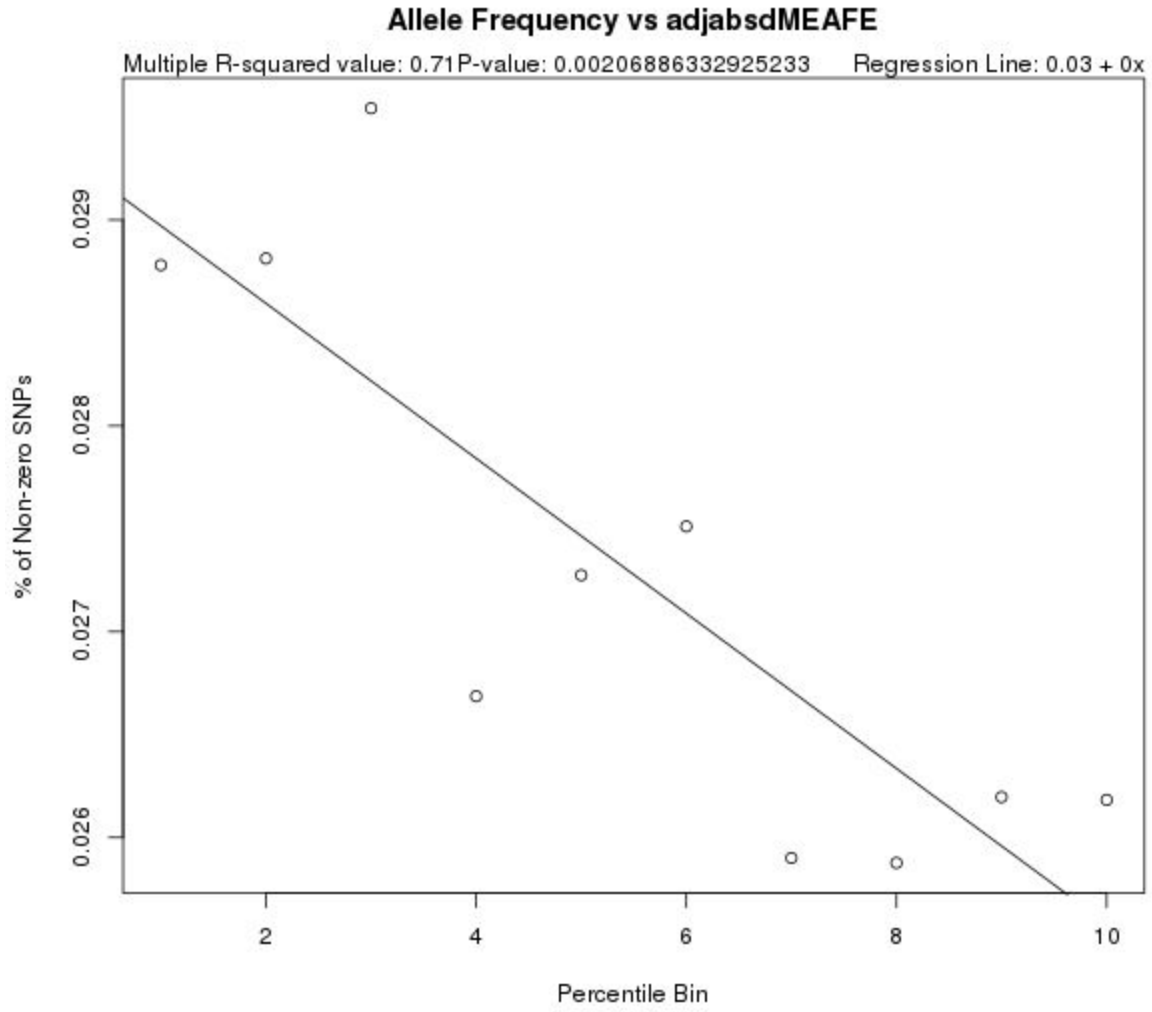
Supplementary Figure 43: Mean/Median GERP Score vs. Binned Change in Ensemble Free Energy (dEFE) for 3' UTR Variants



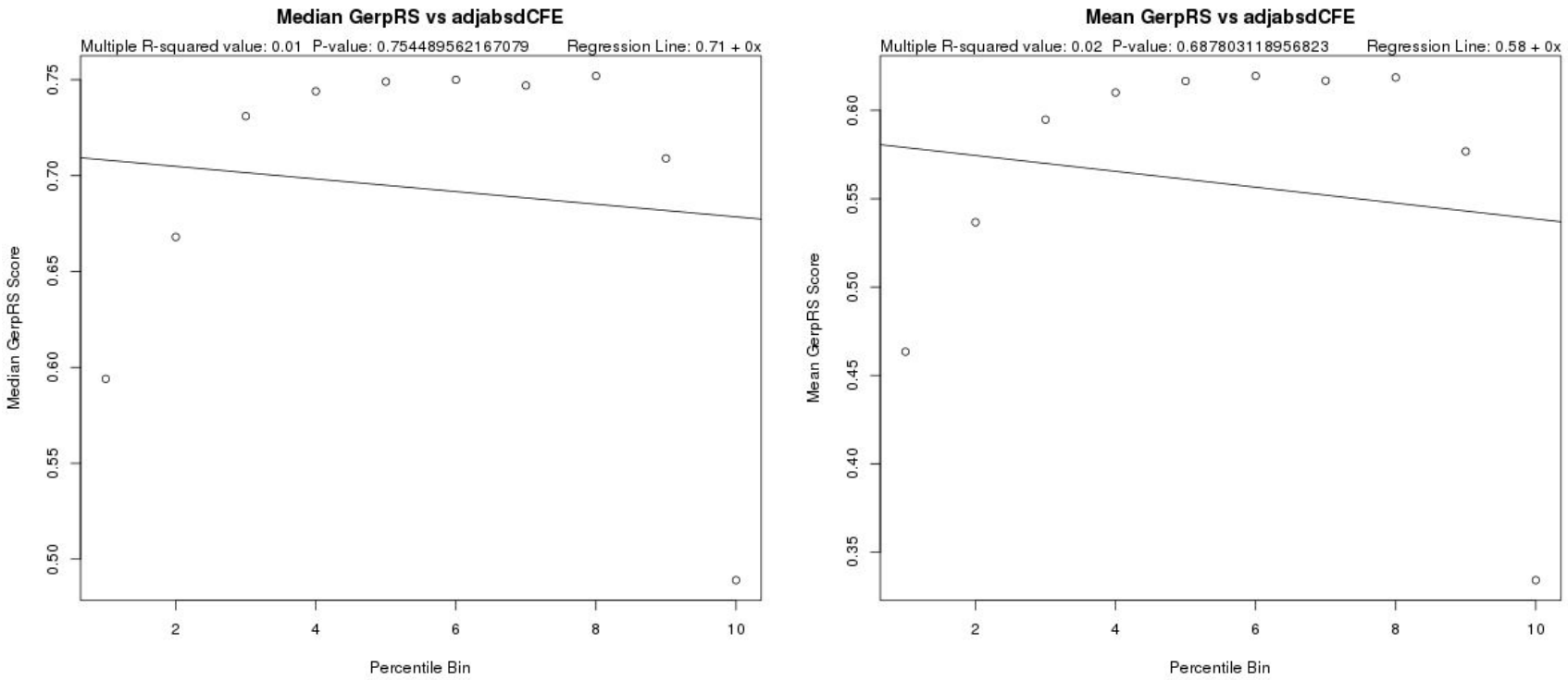
Supplementary Figure 44: % Non-zero Allele Frequency vs. Binned Change in Ensemble Free Energy (dEFE) for 3' UTR Variants



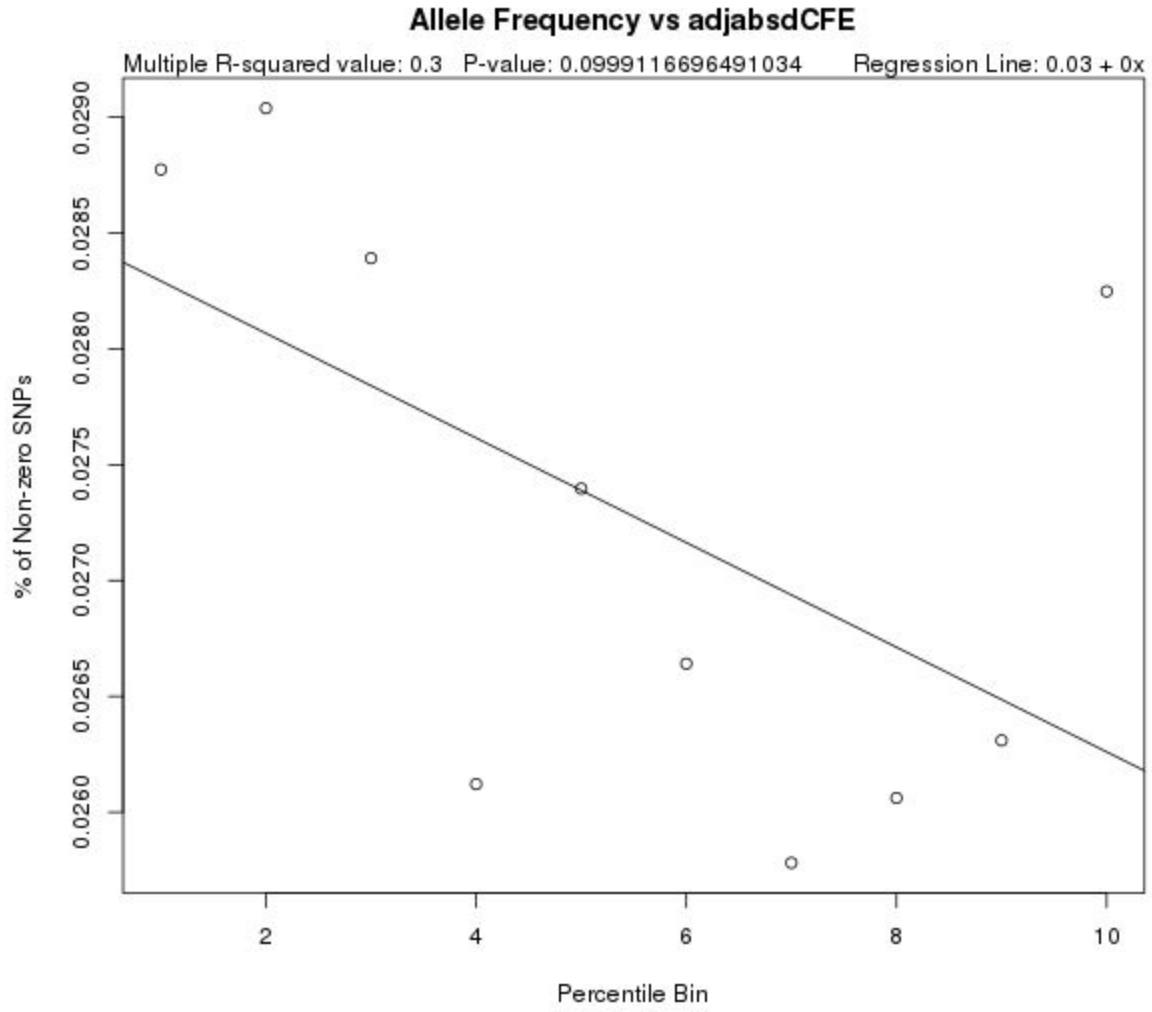
Supplementary Figure 45: Mean/Median GERP Score vs. Binned Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for 3' UTR Variants



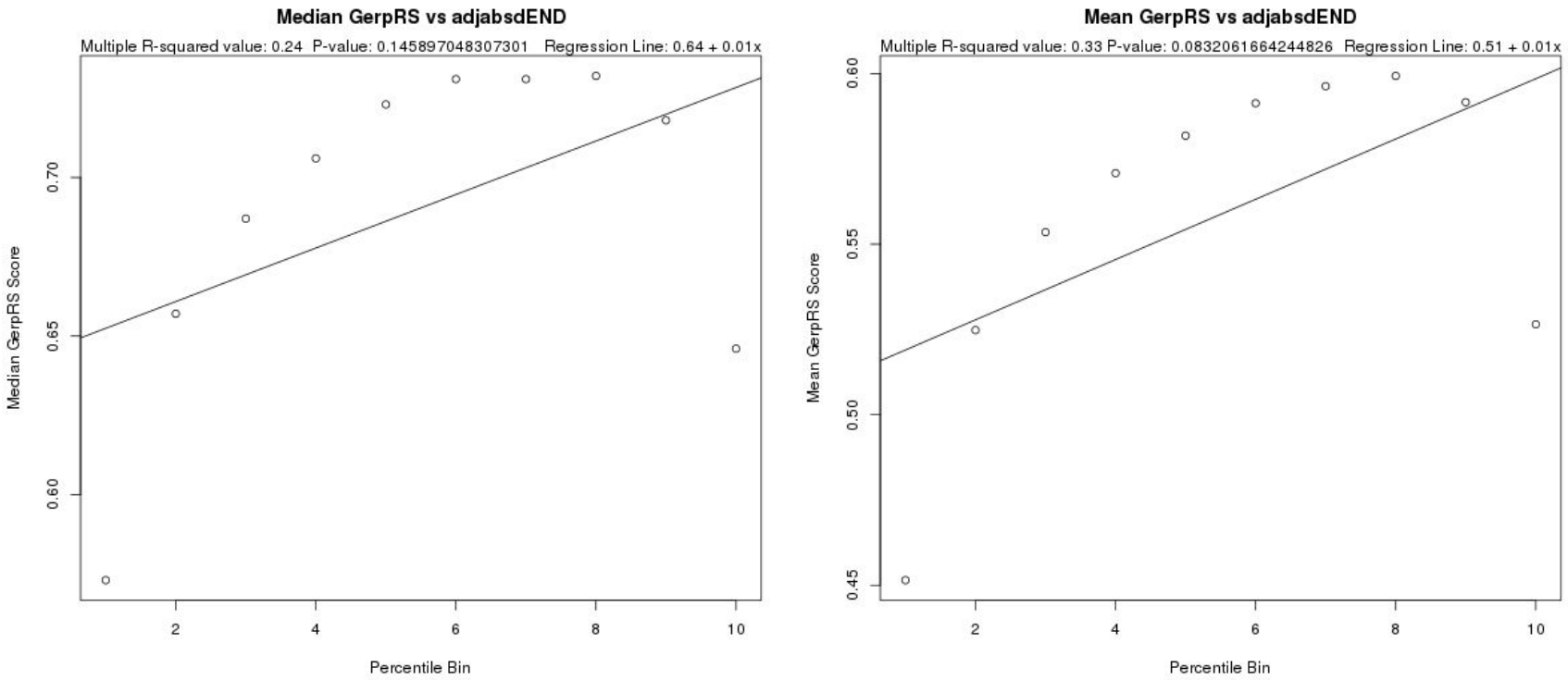
Supplementary Figure 46: % Non-zero Allele Frequency vs. Binned Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for 3' UTR Variants



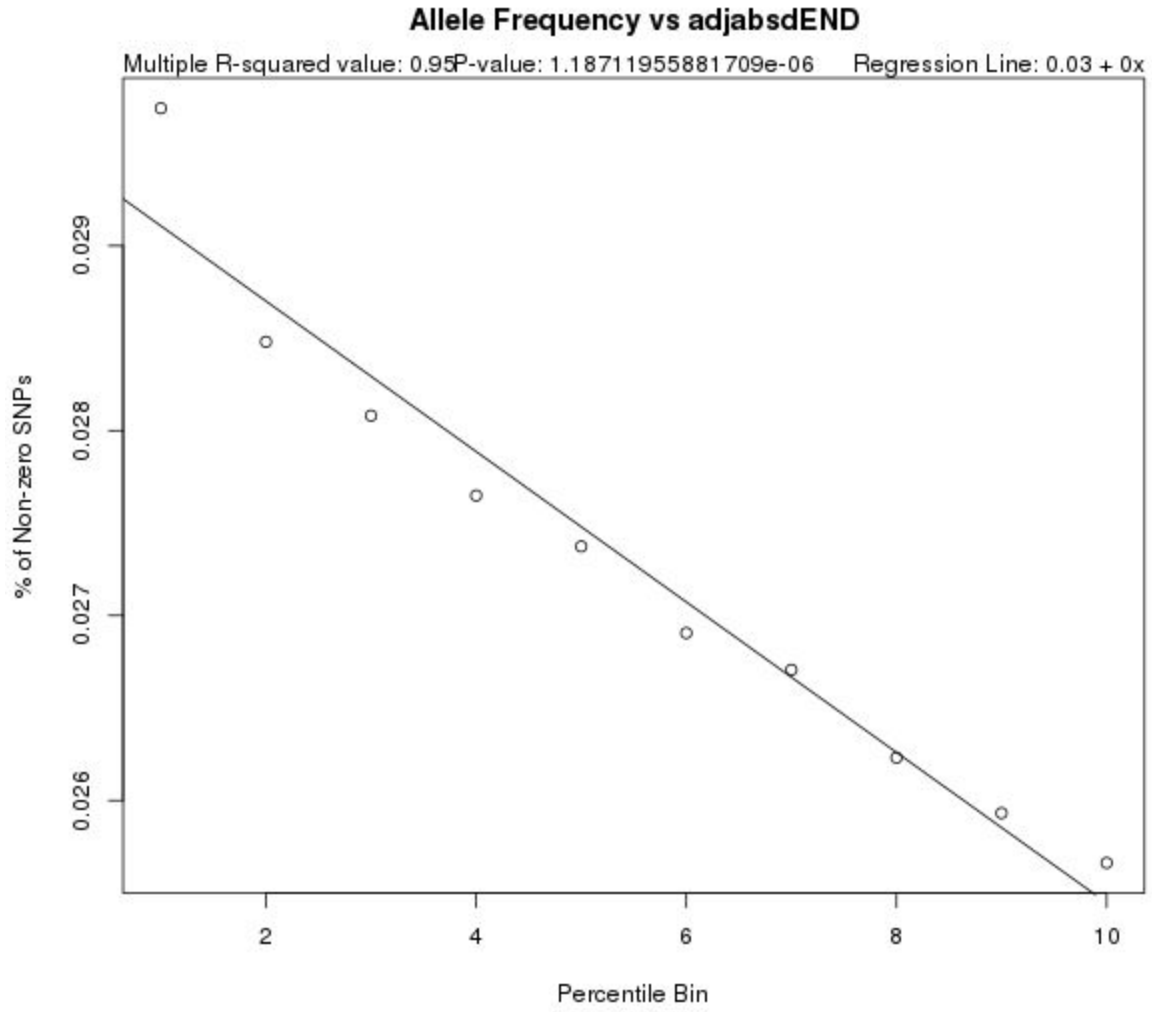
Supplementary Figure 47: Mean/Median GERP Score vs. Binned Change in Free Energy of the Centroid (dCFE) for 3' UTR Variants



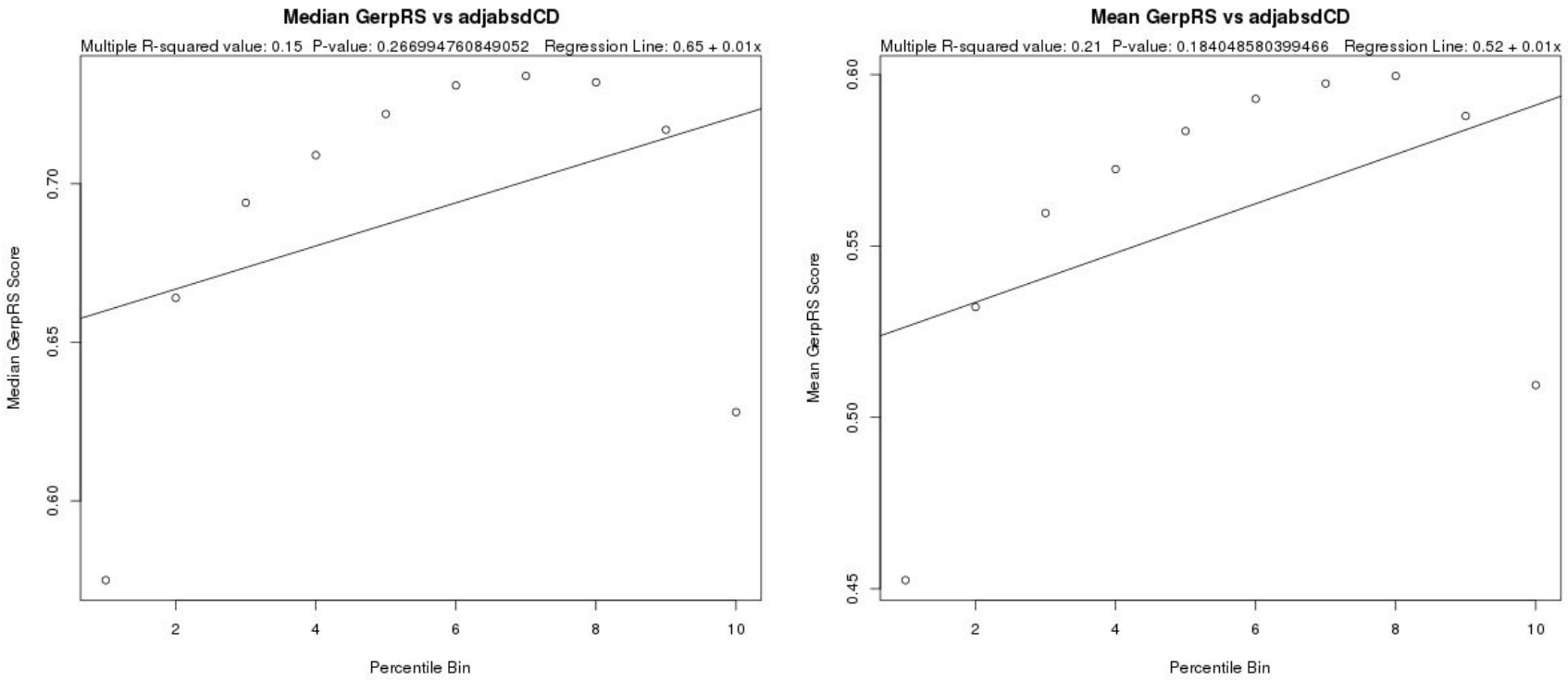
Supplementary Figure 48: % Non-zero Allele Frequency vs. Binned Change in Free Energy of the Centroid (dCFE) for 3' UTR Variants



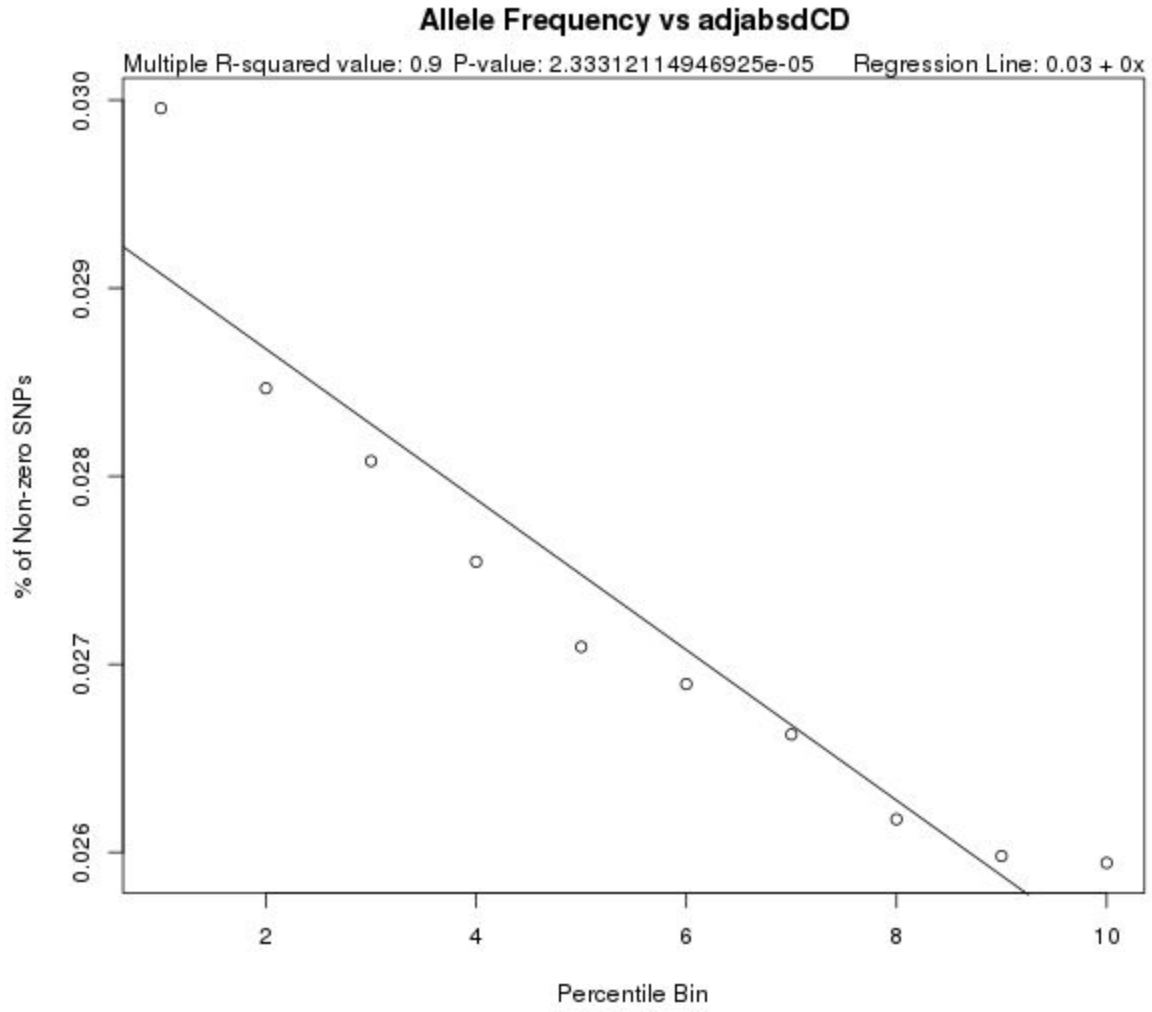
Supplementary Figure 49: Mean/Median GERP Score vs. Binned Change in Ensemble Diversity (dEND) for 3' UTR Variants



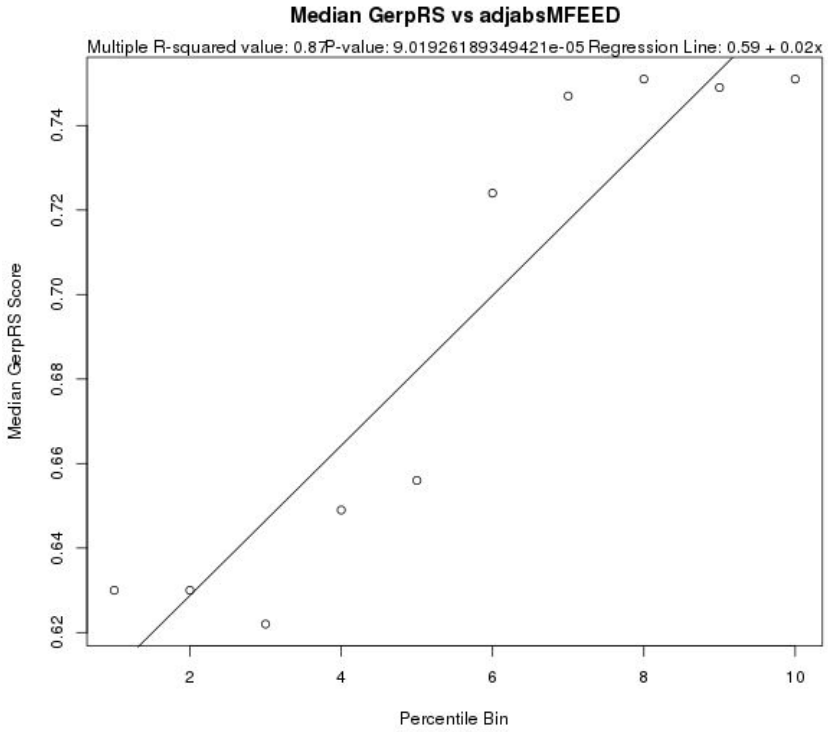
Supplementary Figure 50: % Non-zero Allele Frequency vs. Binned Change in Ensemble Diversity (dEND) for 3' UTR Variants

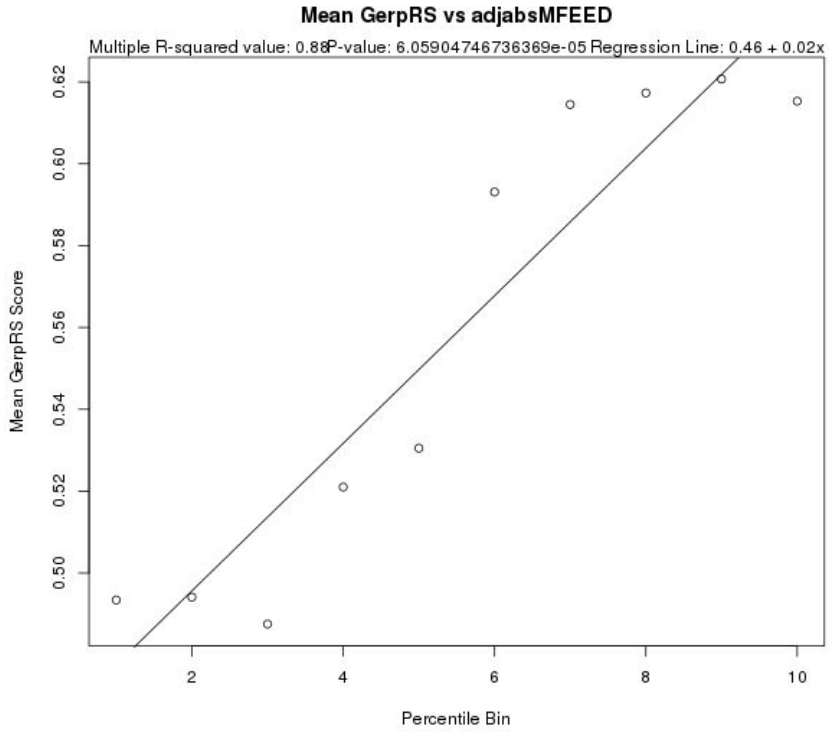


Supplementary Figure 51: Mean/Median GERP Score vs. Binned Change in Distance of the Ensemble of Structures to the Centroid (dCD) for 3' UTR Variants

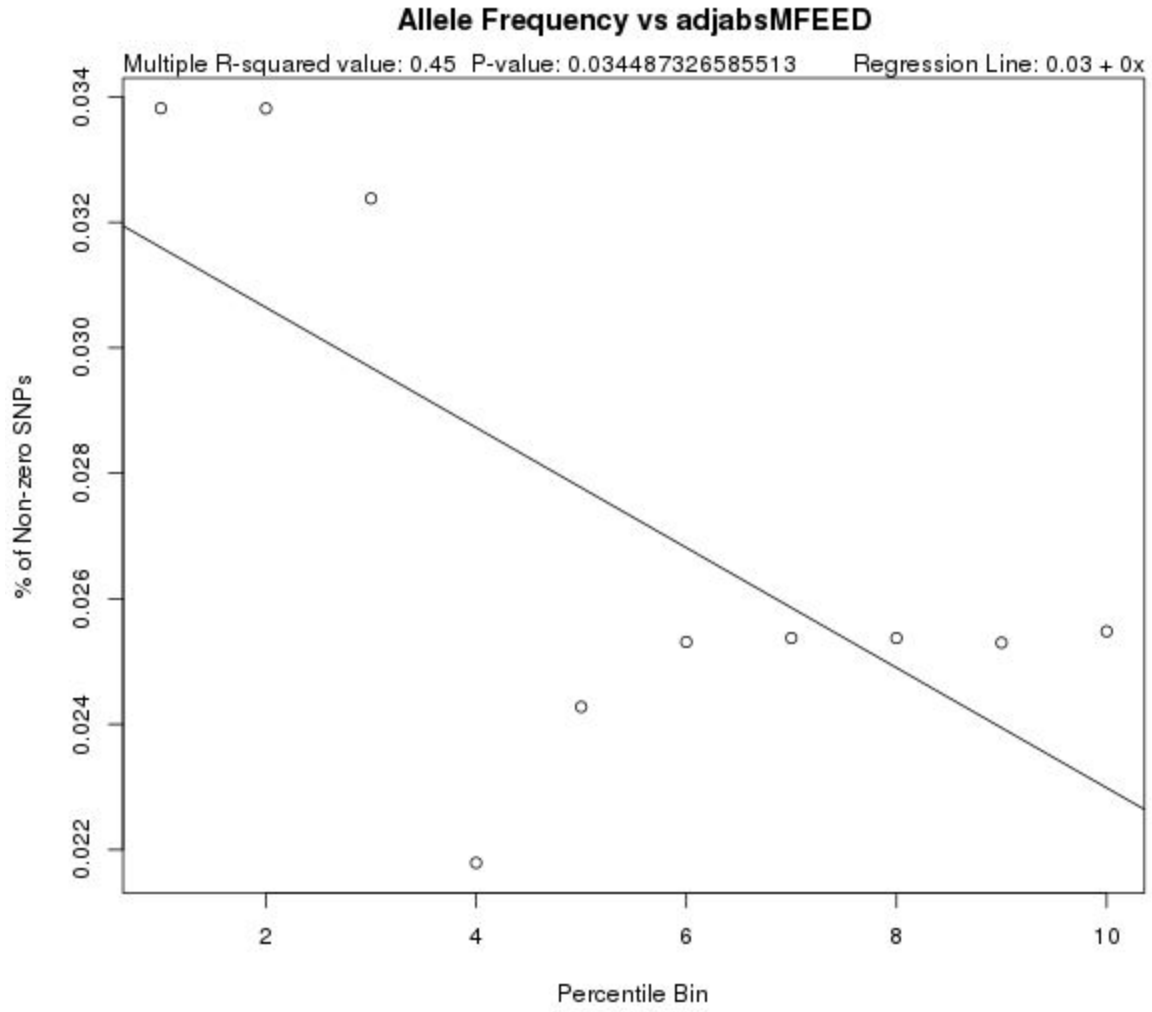


Supplementary Figure 52: % Non-zero Allele Frequency vs. Binned Change in Distance of the Ensemble of Structures to the Centroid (dCD) for 3' UTR Variants

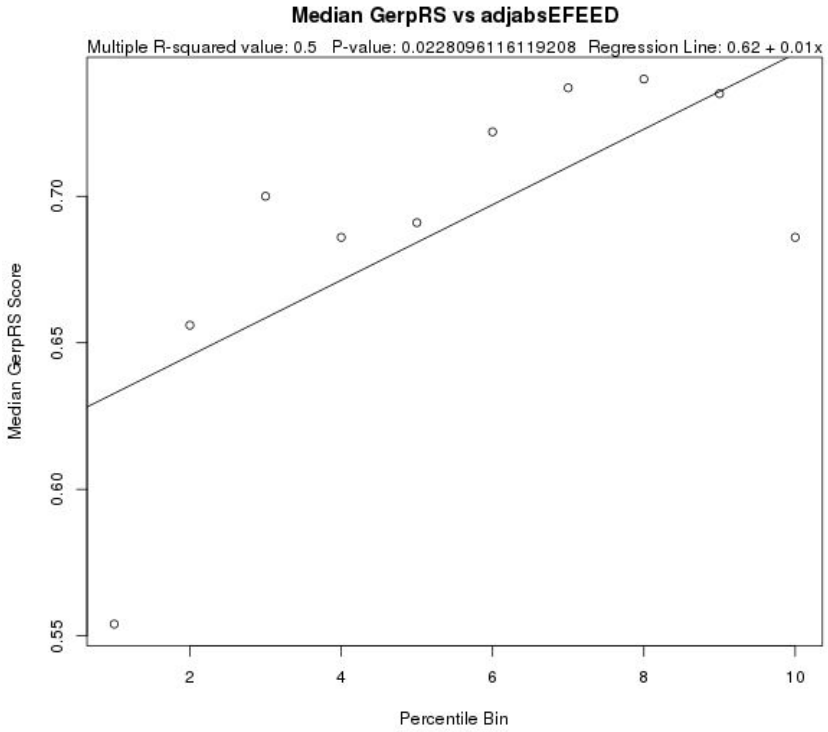


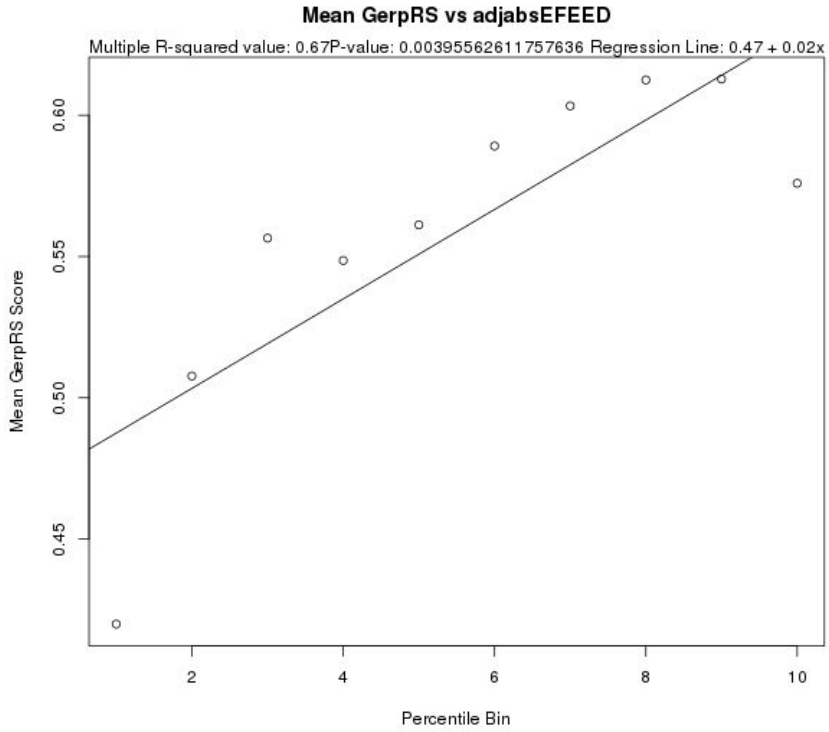


Supplementary Figure 53: Mean/Median GERP Score vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for 3' UTR Variants

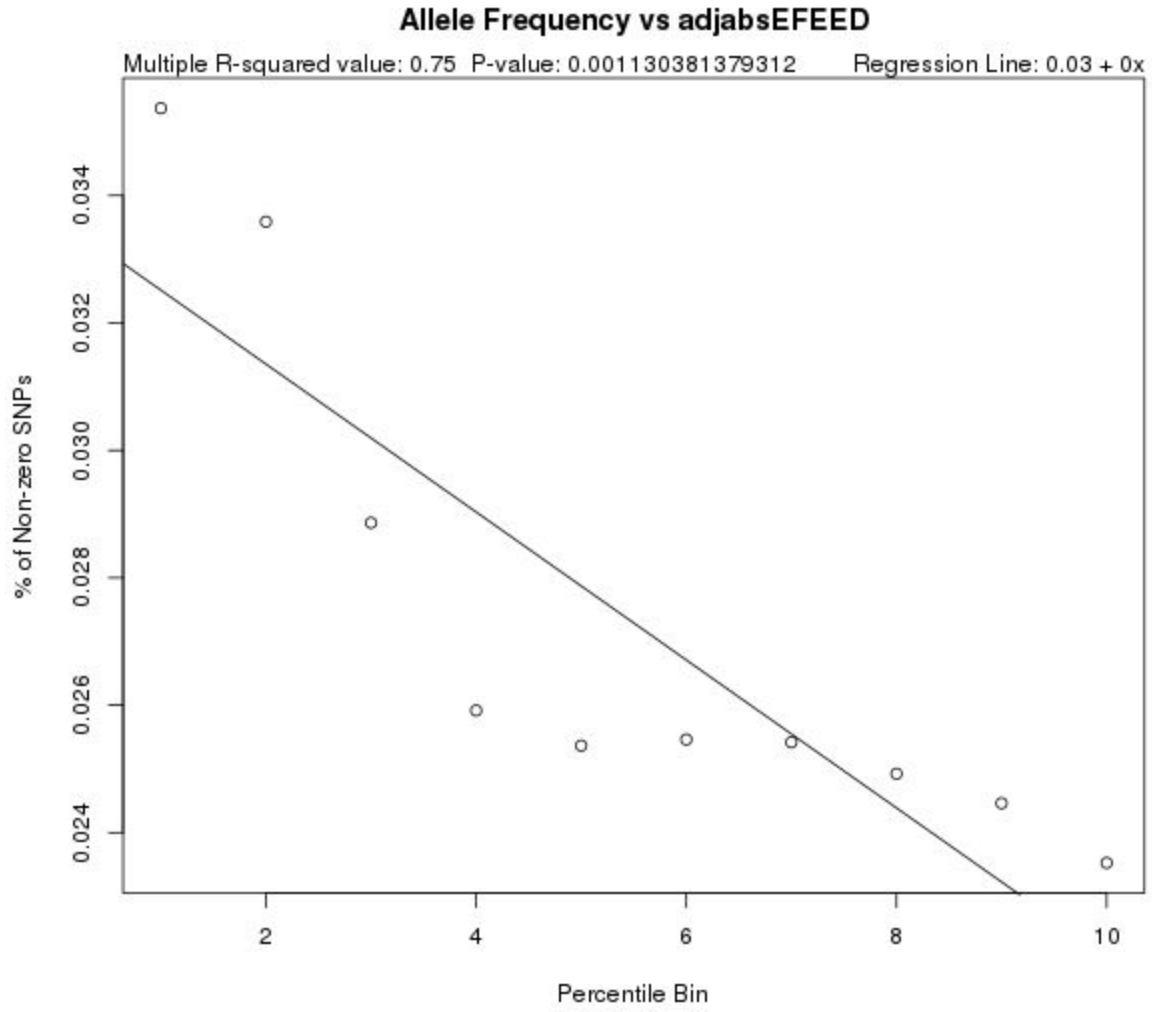


Supplementary Figure 54: % Non-zero Allele Frequency vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for 3' UTR Variants

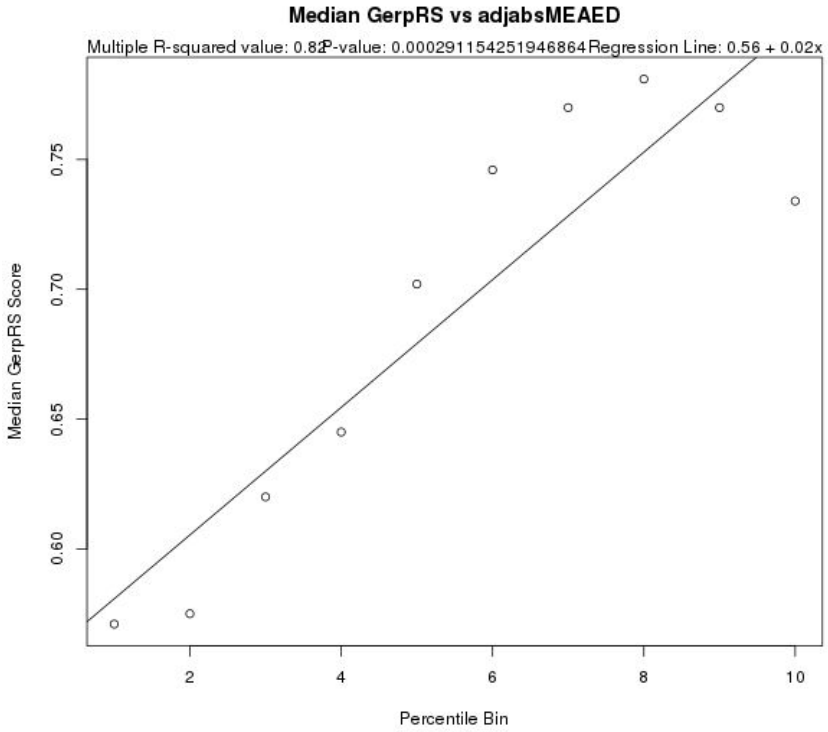


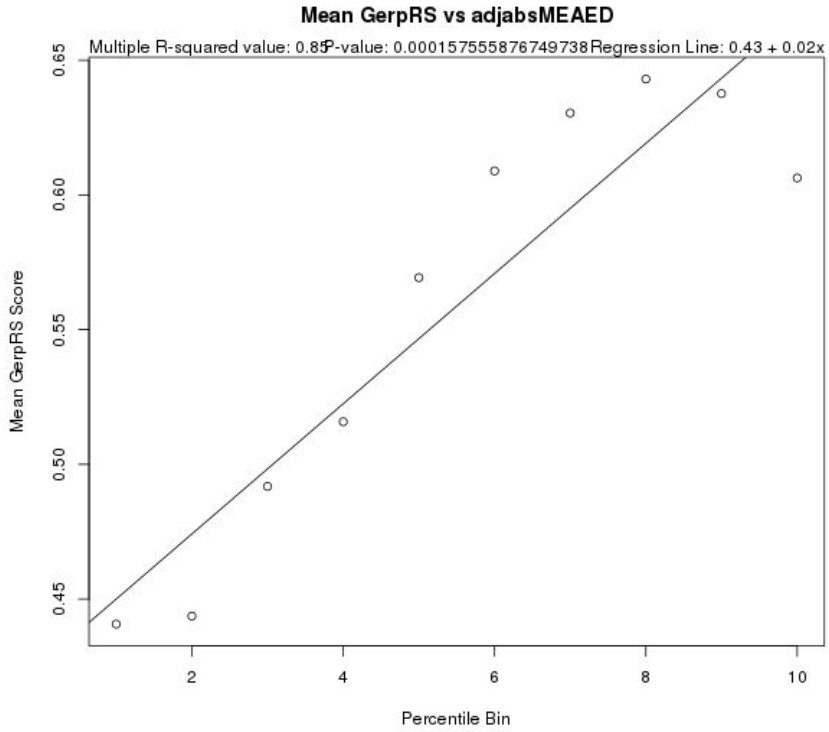


Supplementary Figure 55: Mean/Median GERP Score vs. Edit Distance Between Ensembles (EFEED) for 3' UTR Variants

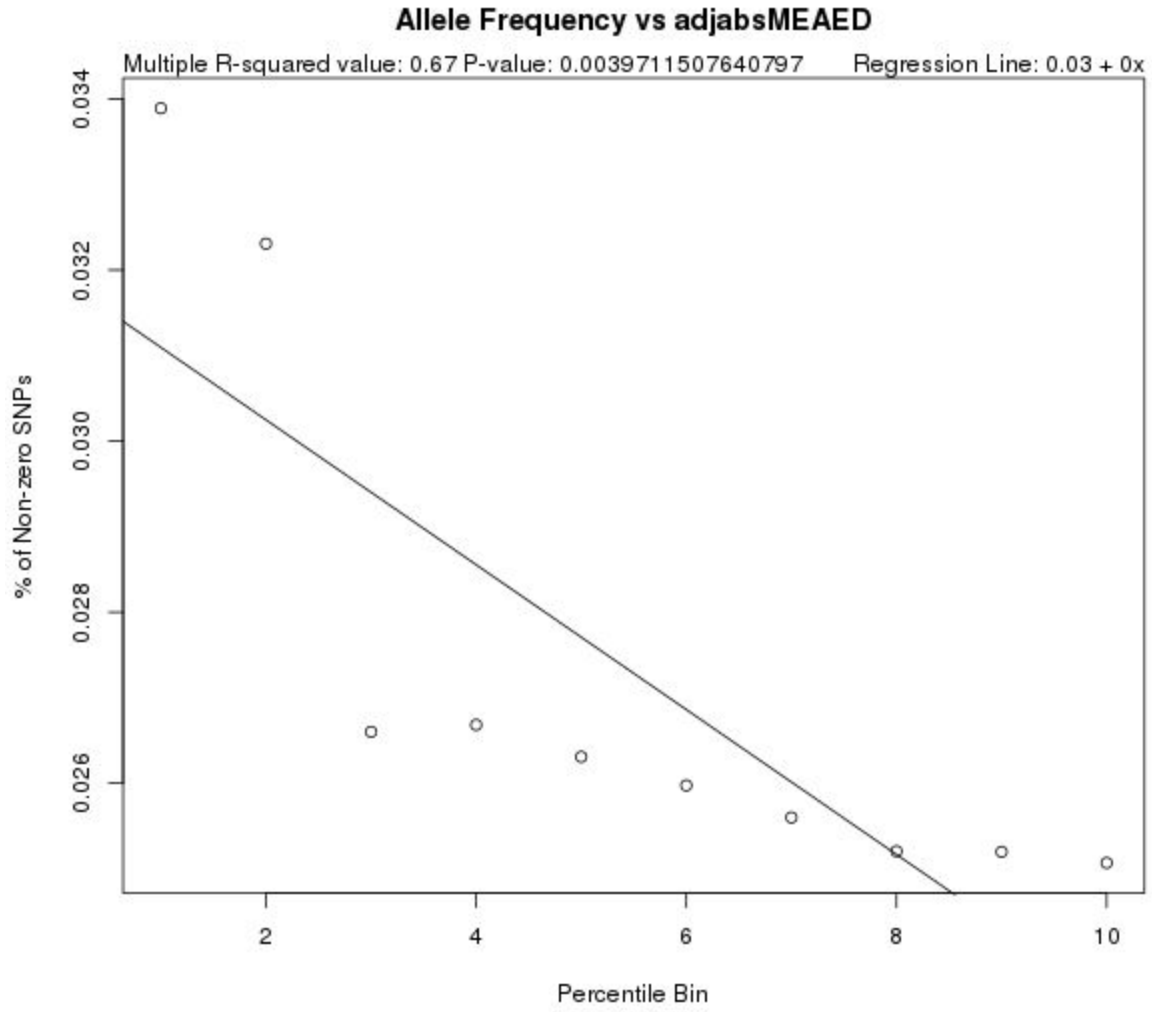


Supplementary Figure 56: % Non-zero Allele Frequency vs. Edit Distance Between Ensembles (EFEED) for 3' UTR Variants

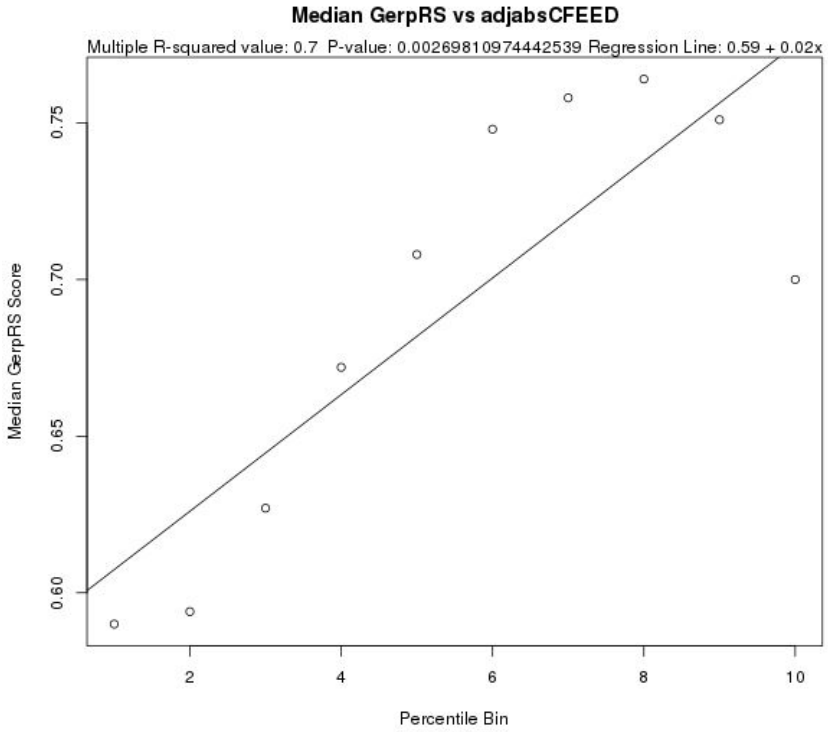


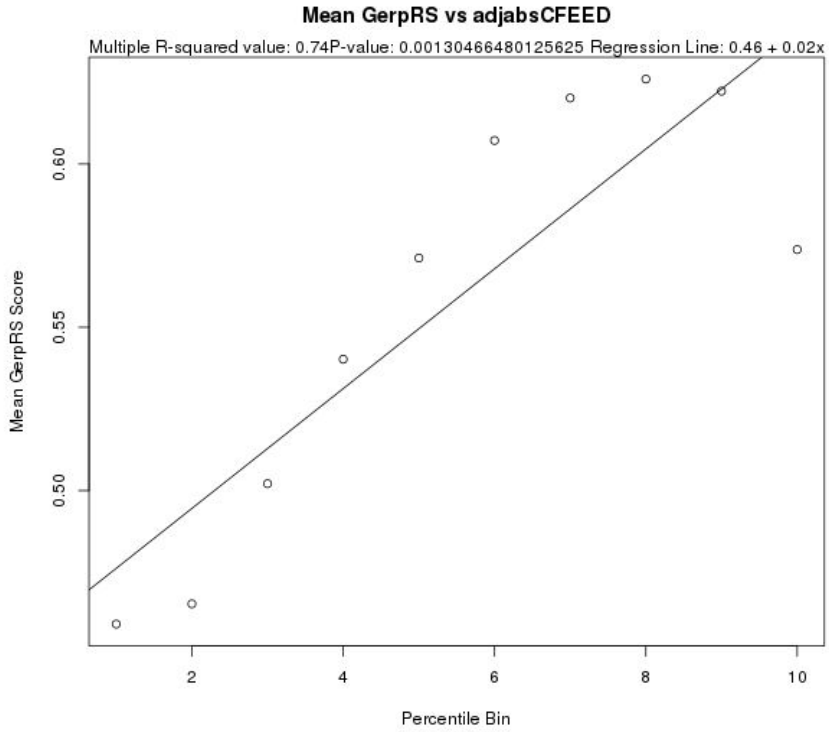


Supplementary Figure 57: Mean/Median GERP Score vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for 3' UTR Variants

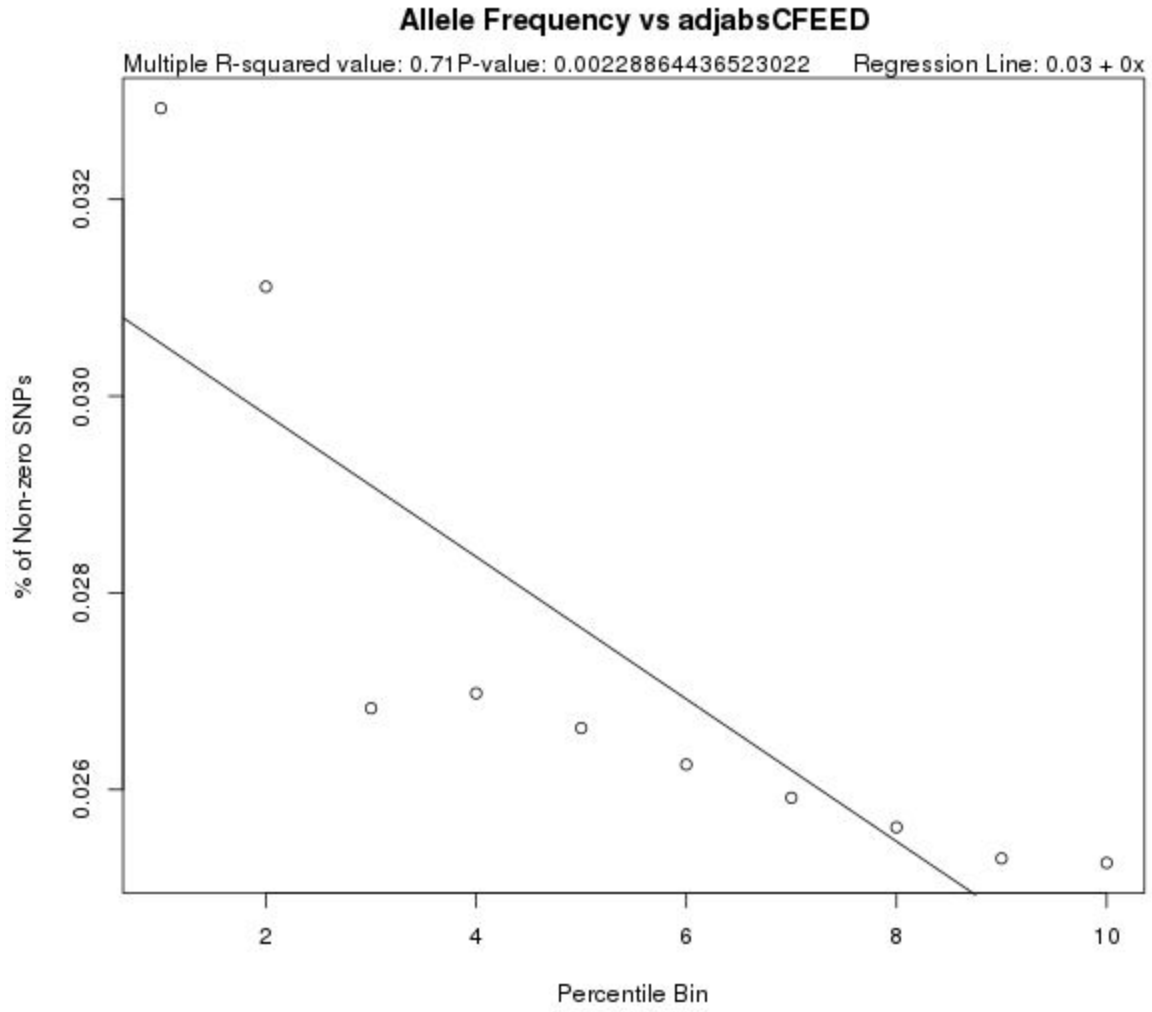


Supplementary Figure 58: % Non-zero Allele Frequency vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for 3' UTR Variants

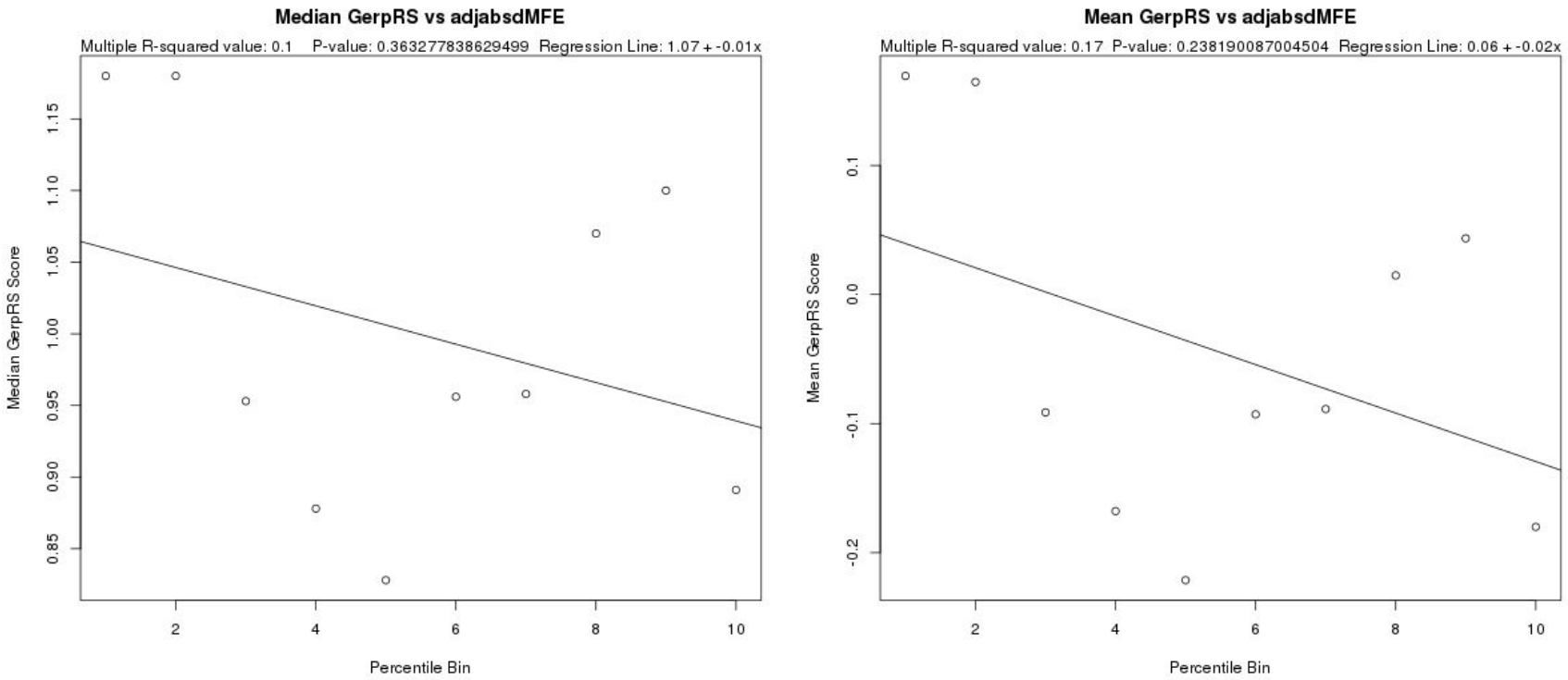




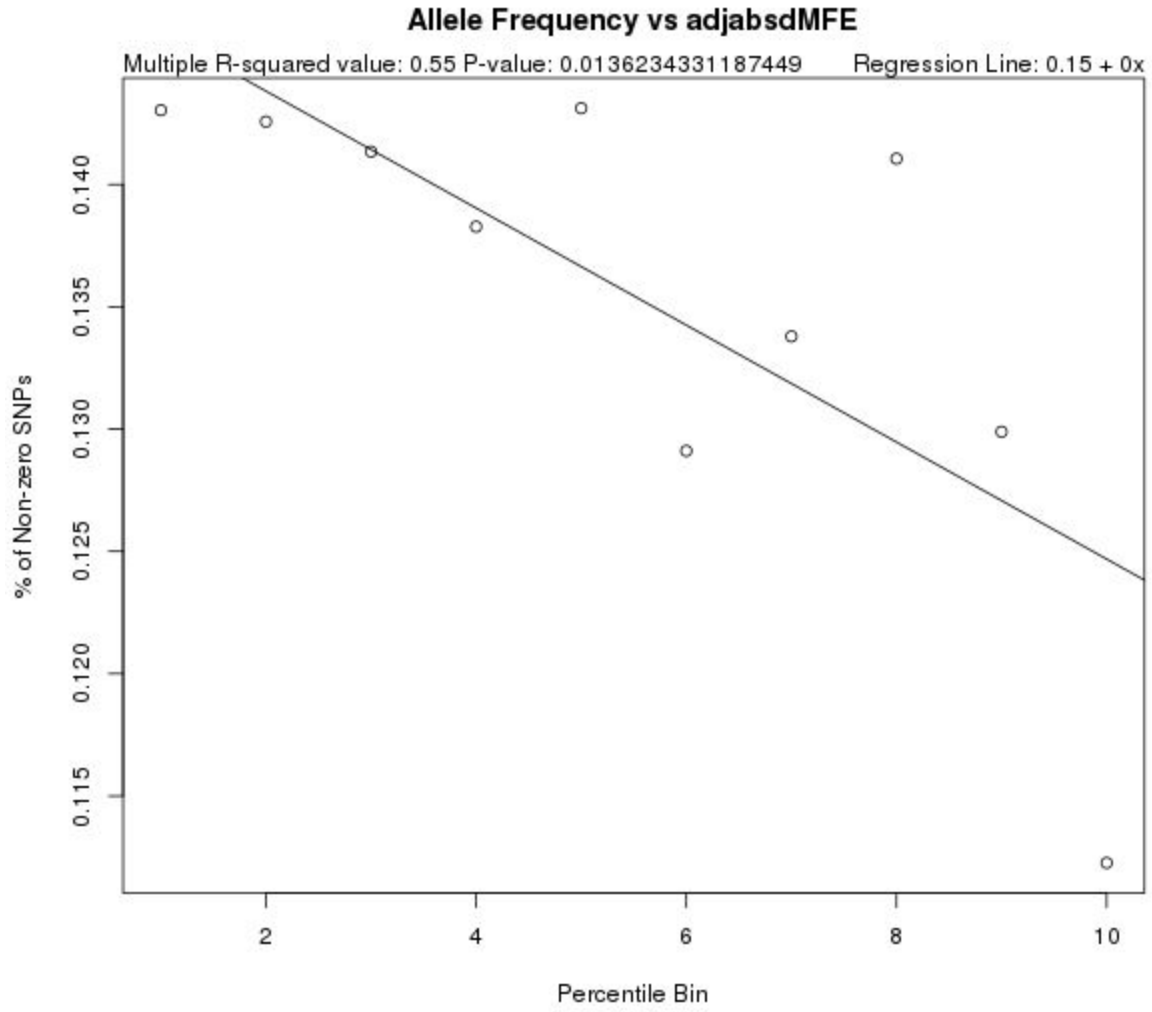
Supplementary Figure 59: Mean/Median GERP Score vs. Edit Distance Between Centroid Structures (CFEED) for 3' UTR Variants



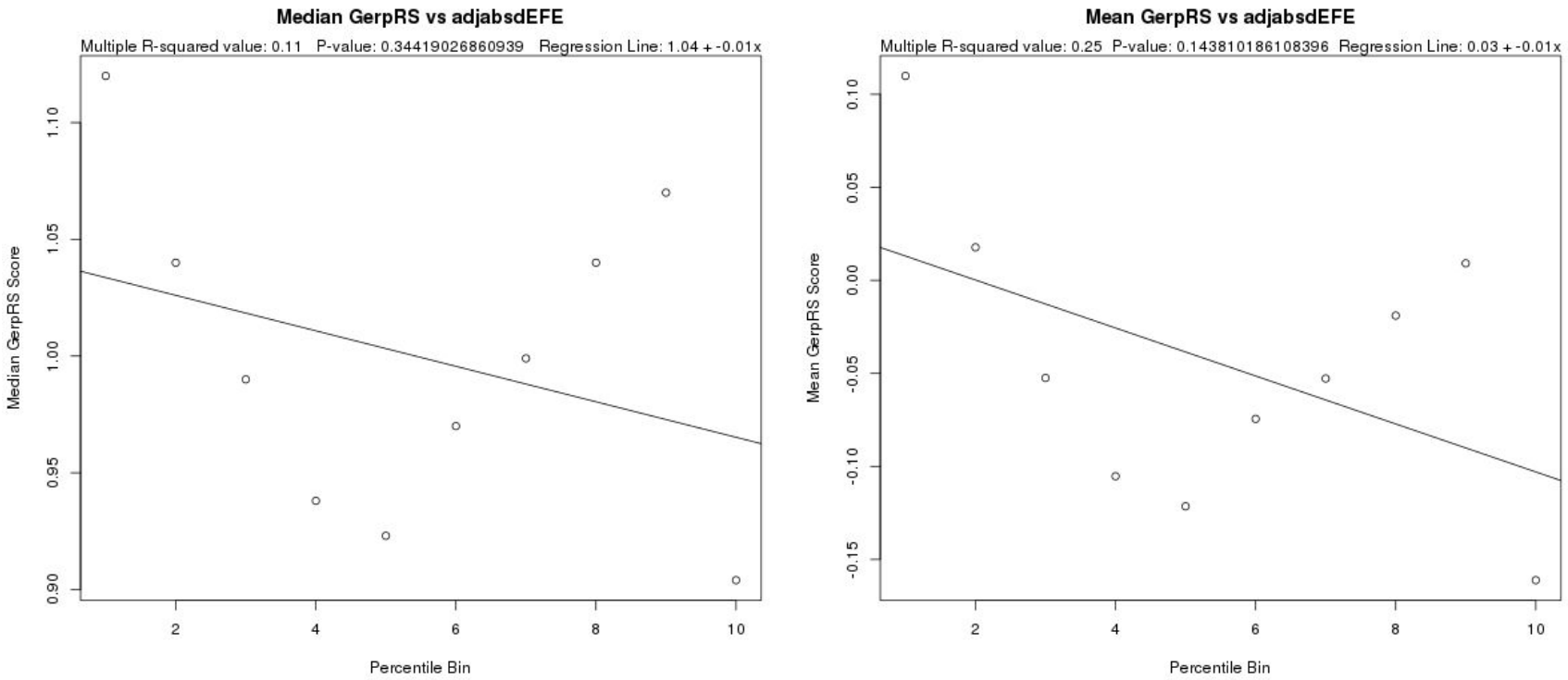
Supplementary Figure 60: % Non-zero Allele Frequency vs. Edit Distance Between Centroid Structures (CFEED) for 3' UTR Variants



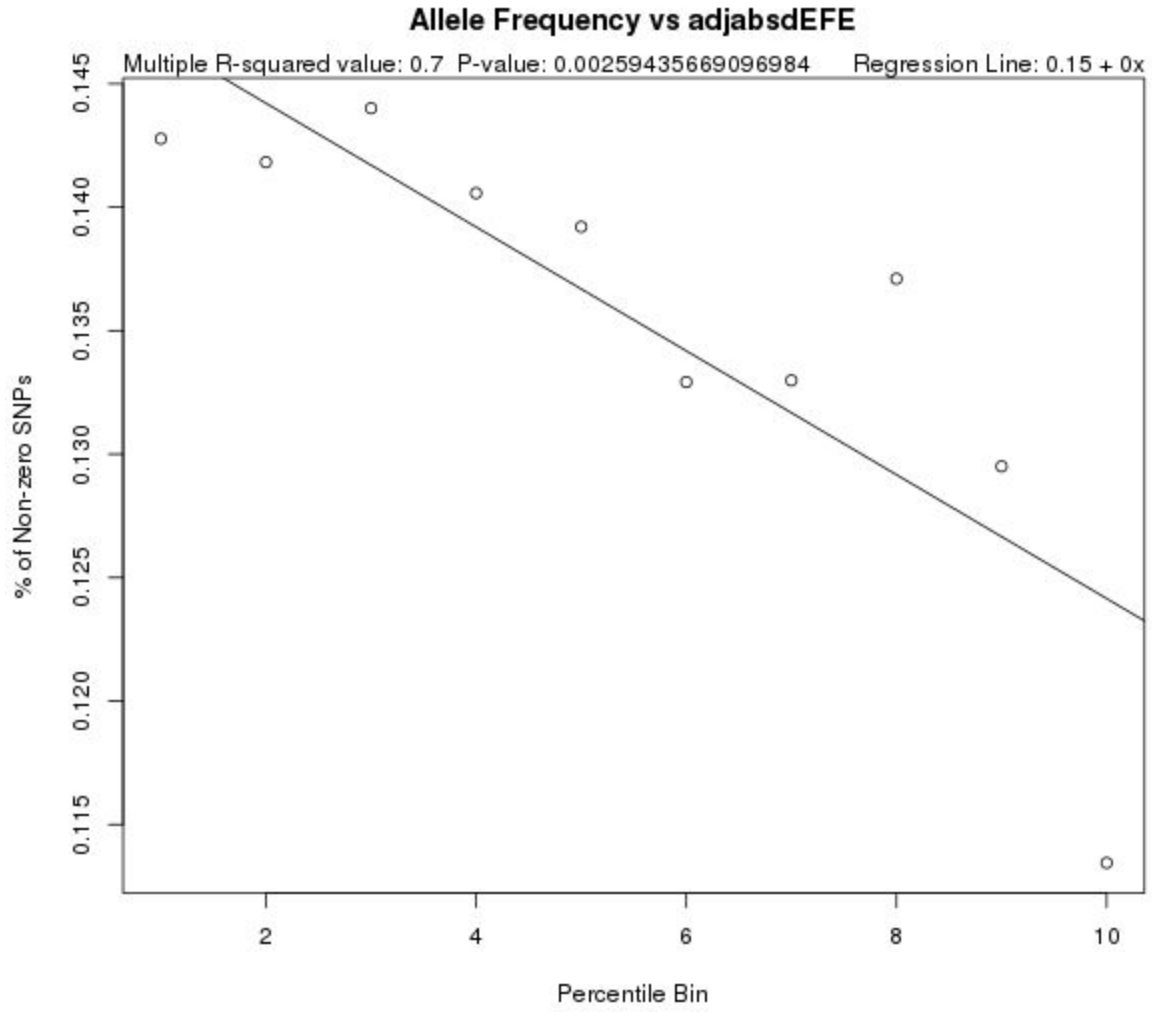
Supplementary Figure 61: Mean/Median GERP Score vs. Binned Change in Minimum Free Energy (dMFE) for Synonymous Variants



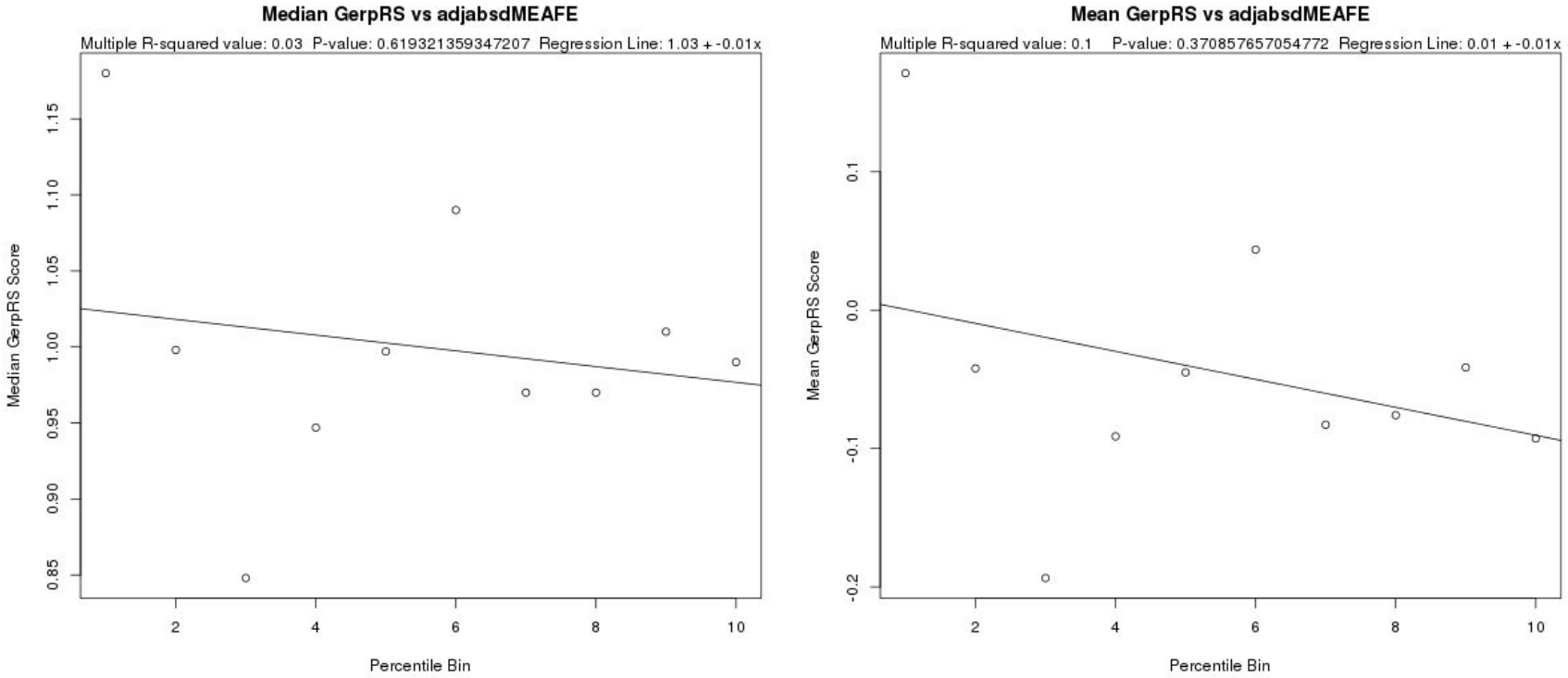
Supplementary Figure 62: % Non-zero Allele Frequency vs. Binned Change in Minimum Free Energy (dMFE) for Synonymous Variants



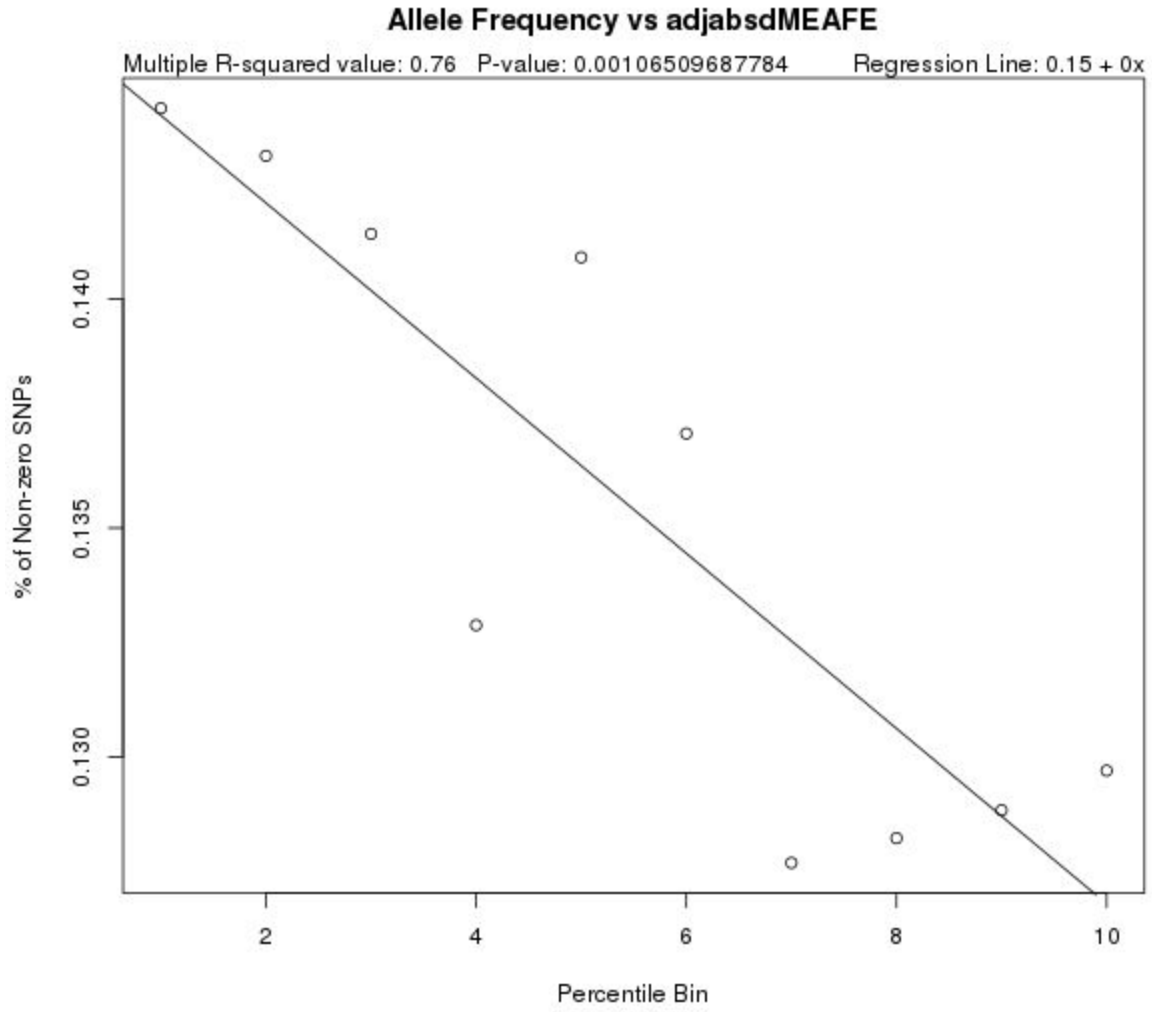
Supplementary Figure 63: Mean/Median GERP Score vs. Binned Change in Ensemble Free Energy (dEFE) for Synonymous Variants



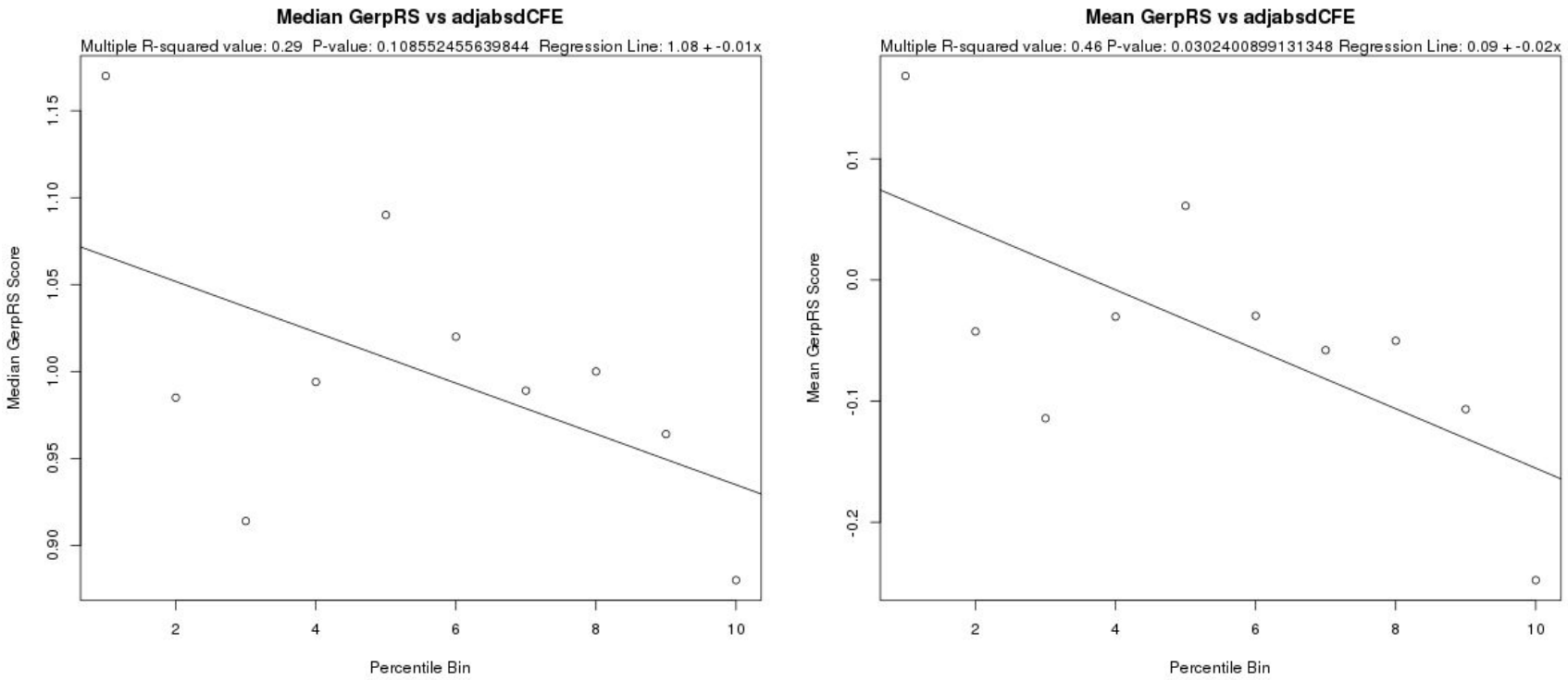
Supplementary Figure 64: % Non-zero Allele Frequency vs. Binned Change in Ensemble Free Energy (dEFE) for Synonymous Variants



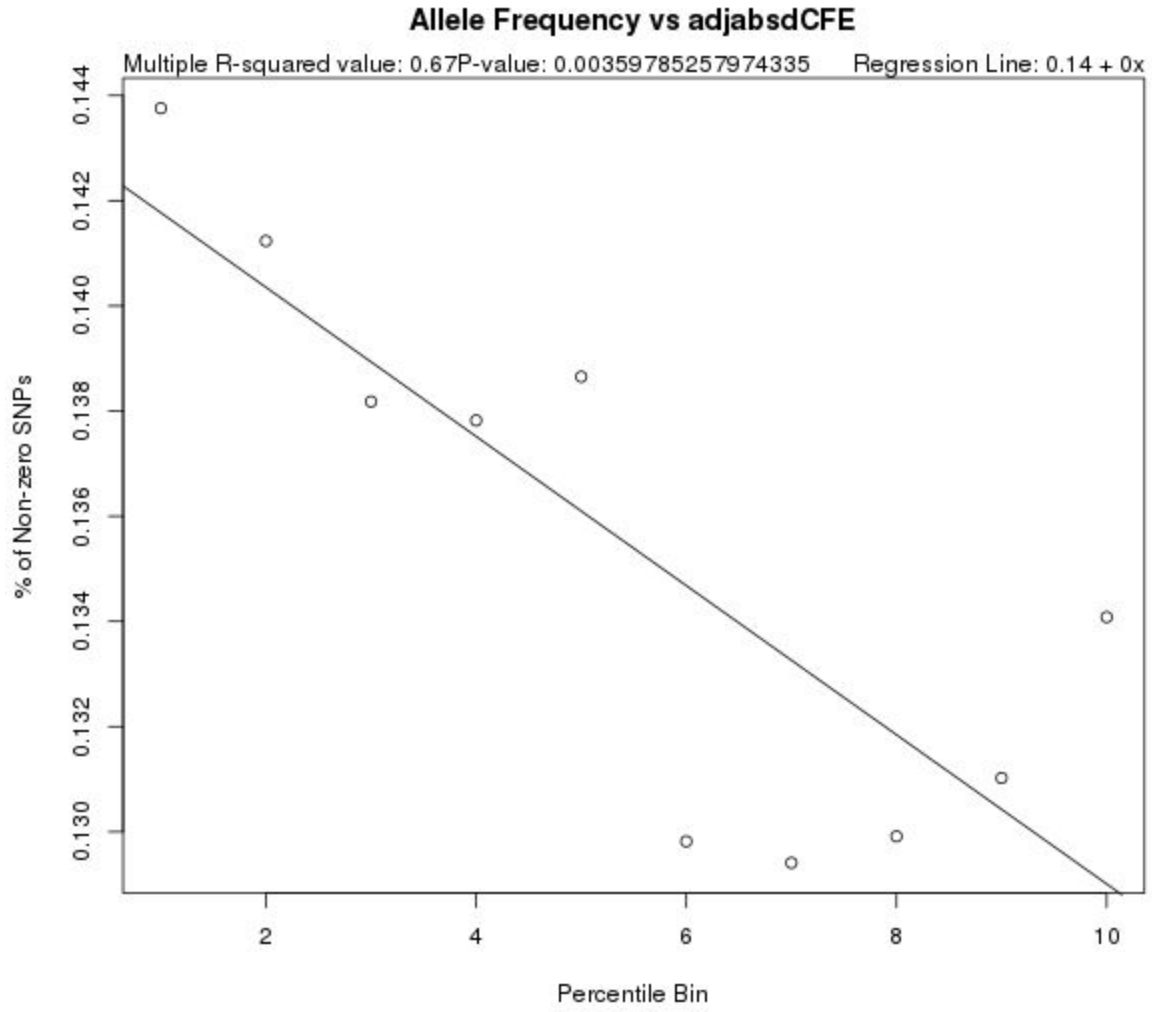
Supplementary Figure 65: Mean/Median GERP Score vs. Binned Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for Synonymous Variants



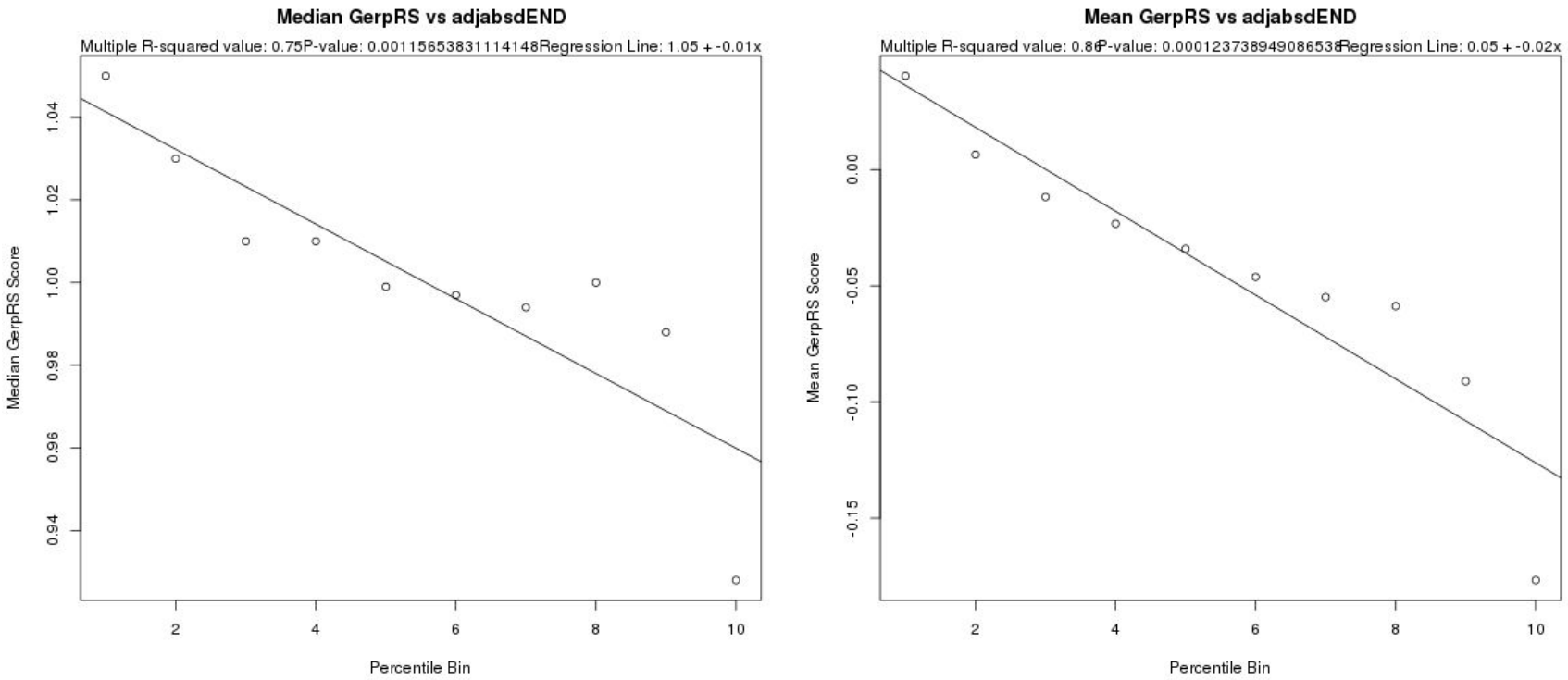
Supplementary Figure 66: % Non-zero Allele Frequency vs. Binned Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for Synonymous Variants



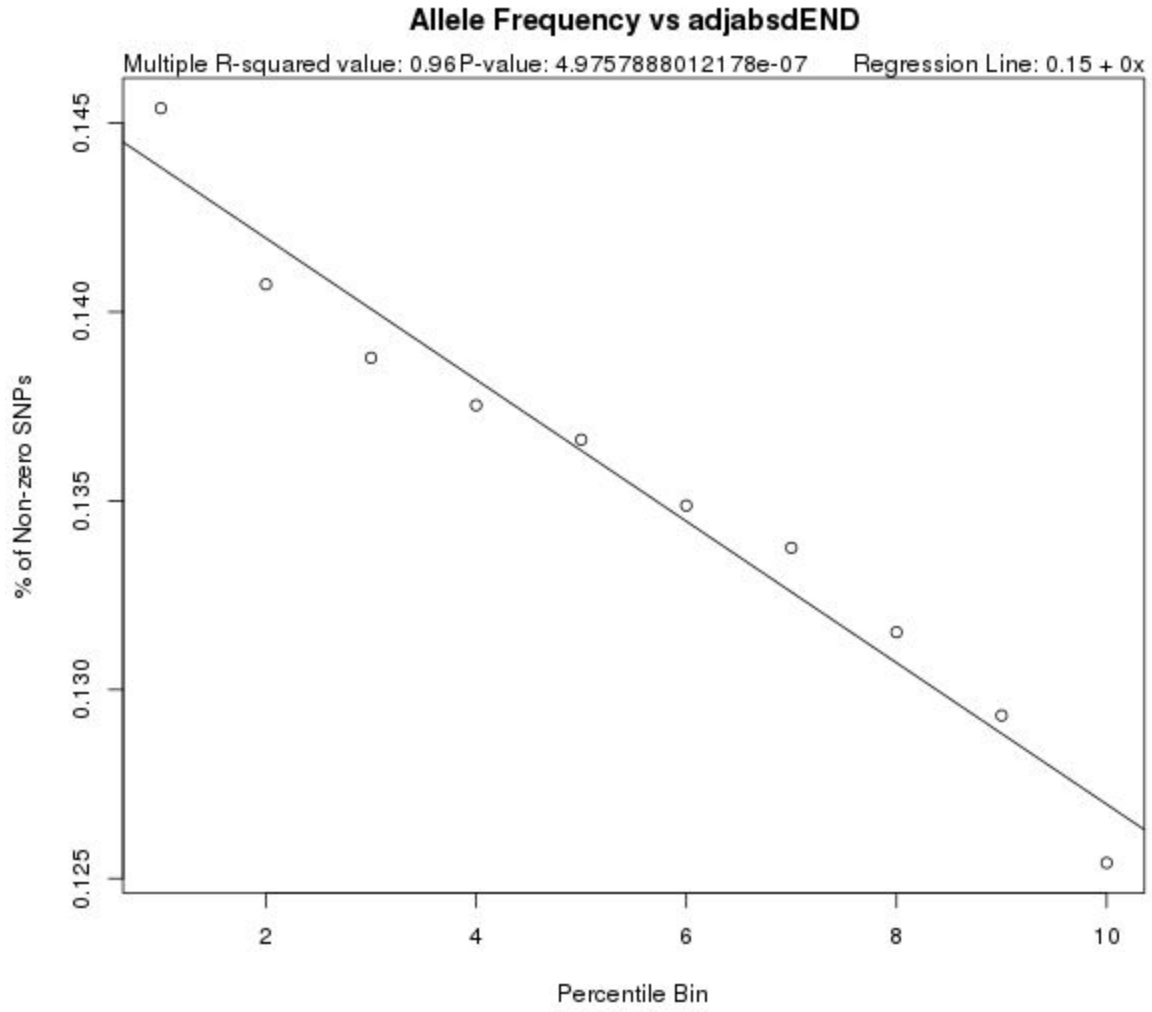
Supplementary Figure 67: Mean/Median GERP Score vs. Binned Change in Free Energy of the Centroid (dCFE) for Synonymous Variants



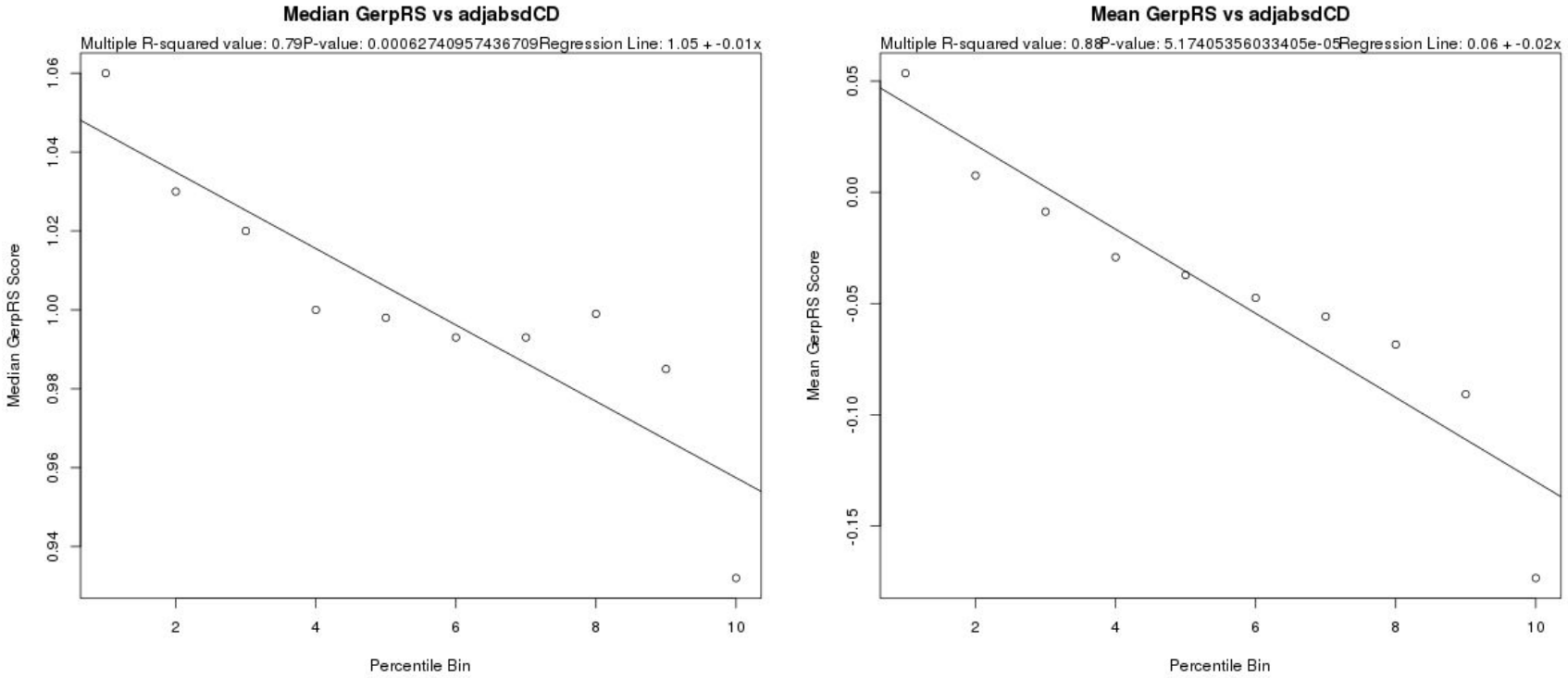
Supplementary Figure 68: % Non-zero Allele Frequency vs. Binned Change in Free Energy of the Centroid (dCFE) for Synonymous Variants



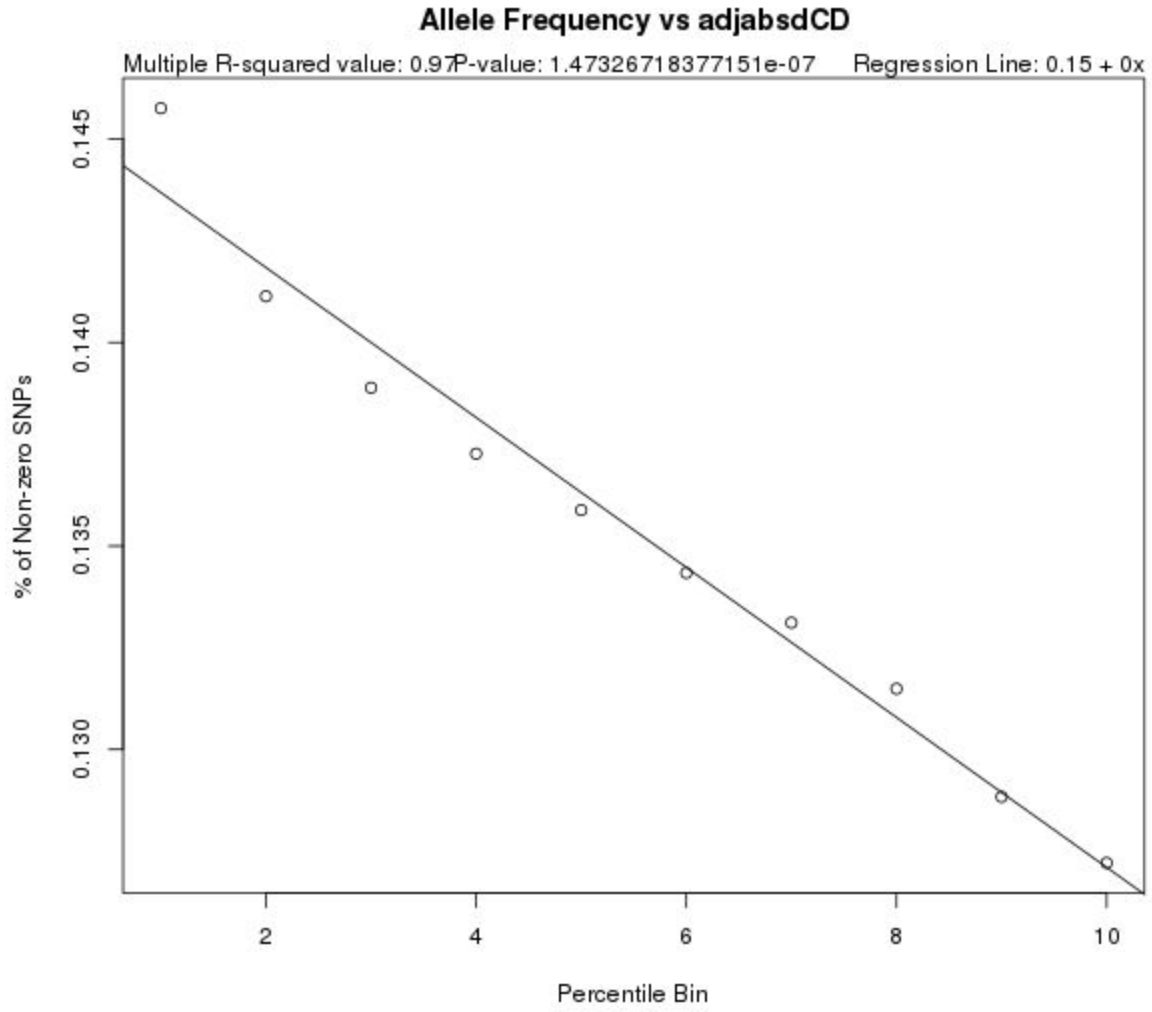
Supplementary Figure 69: Mean/Median GERP Score vs. Binned Change in Ensemble Diversity (dEND) for Synonymous Variants



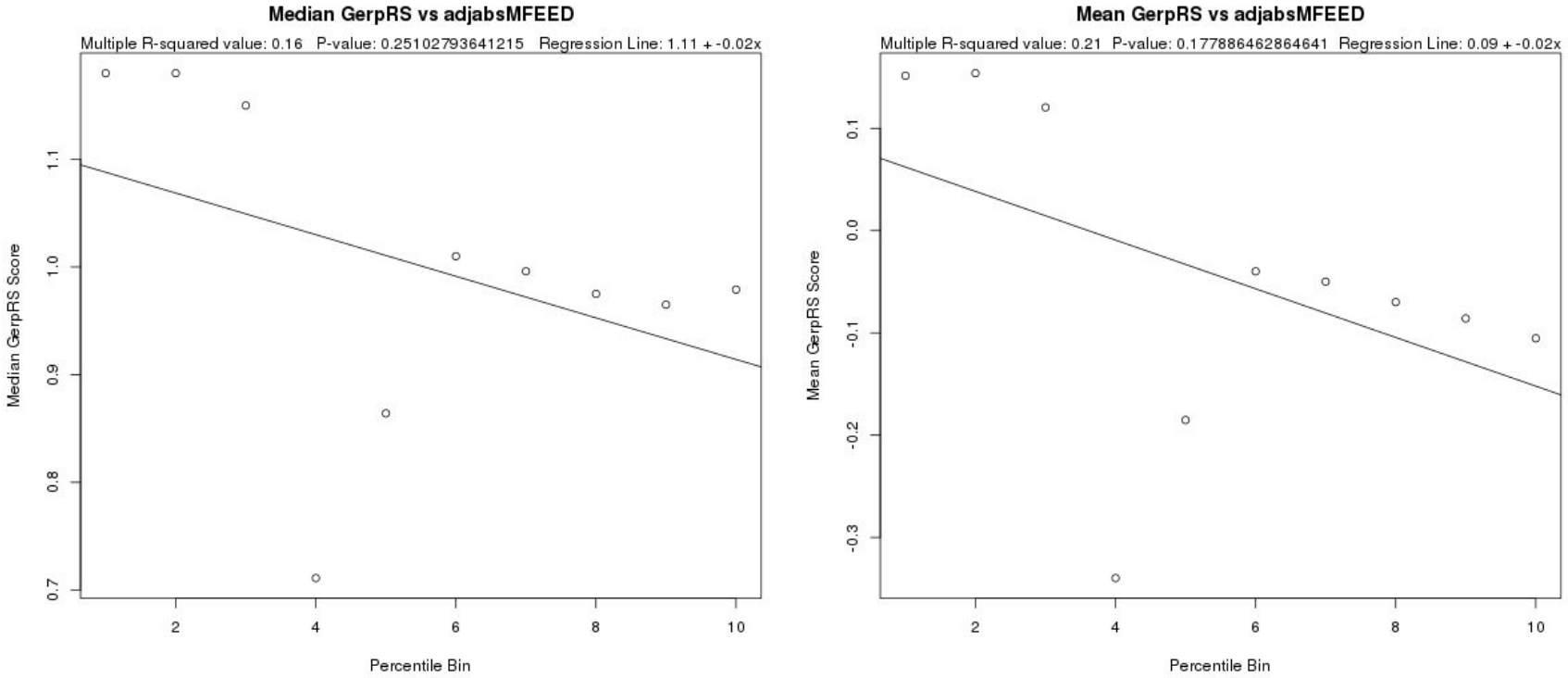
Supplementary Figure 70: % Non-zero Allele Frequency vs. Binned Change in Ensemble Diversity (dEND) for Synonymous Variants



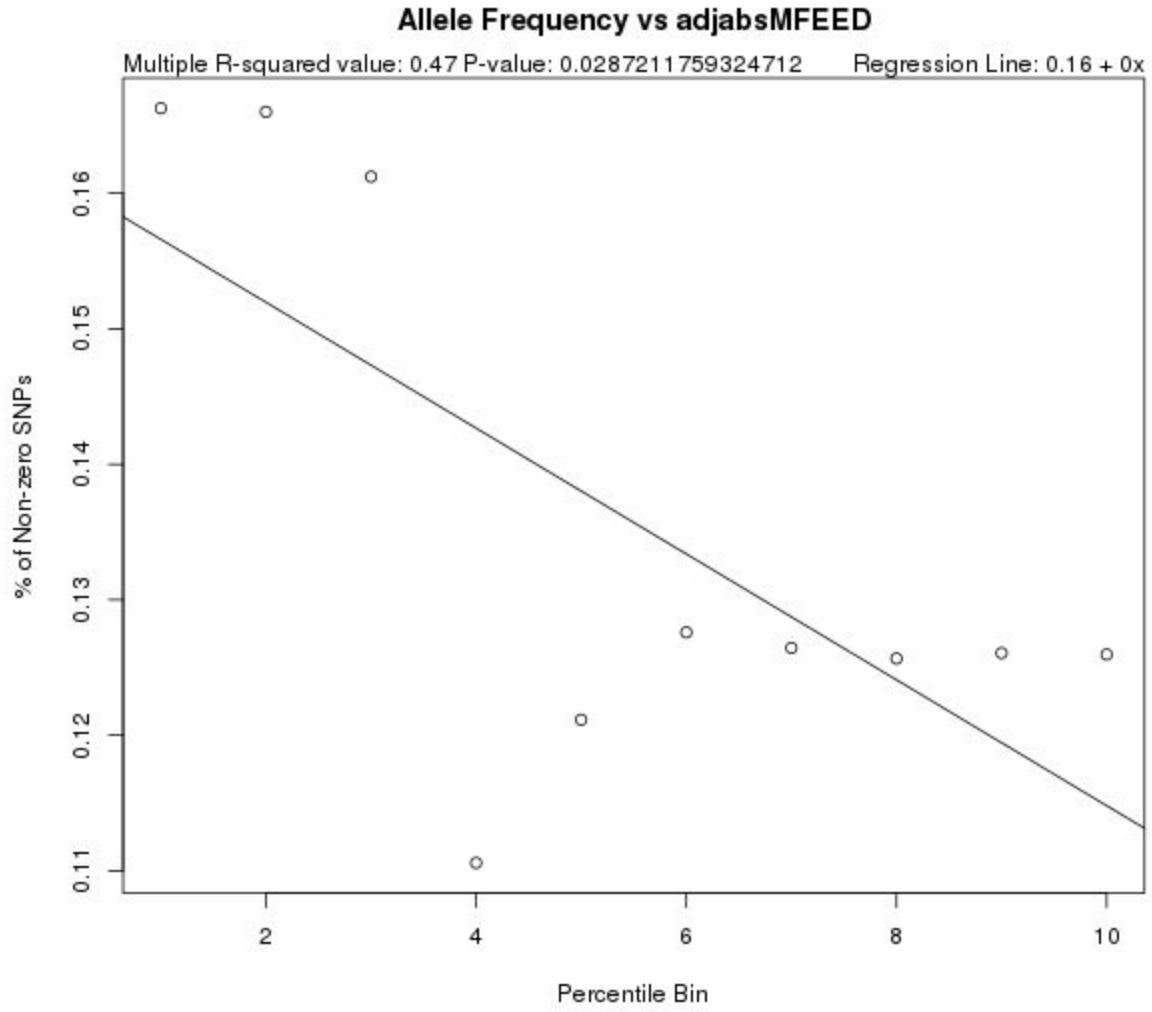
Supplementary Figure 71: Mean/Median GERP Score vs. Binned Change in Distance of the Ensemble of Structures to the Centroid (dCD) for Synonymous Variants



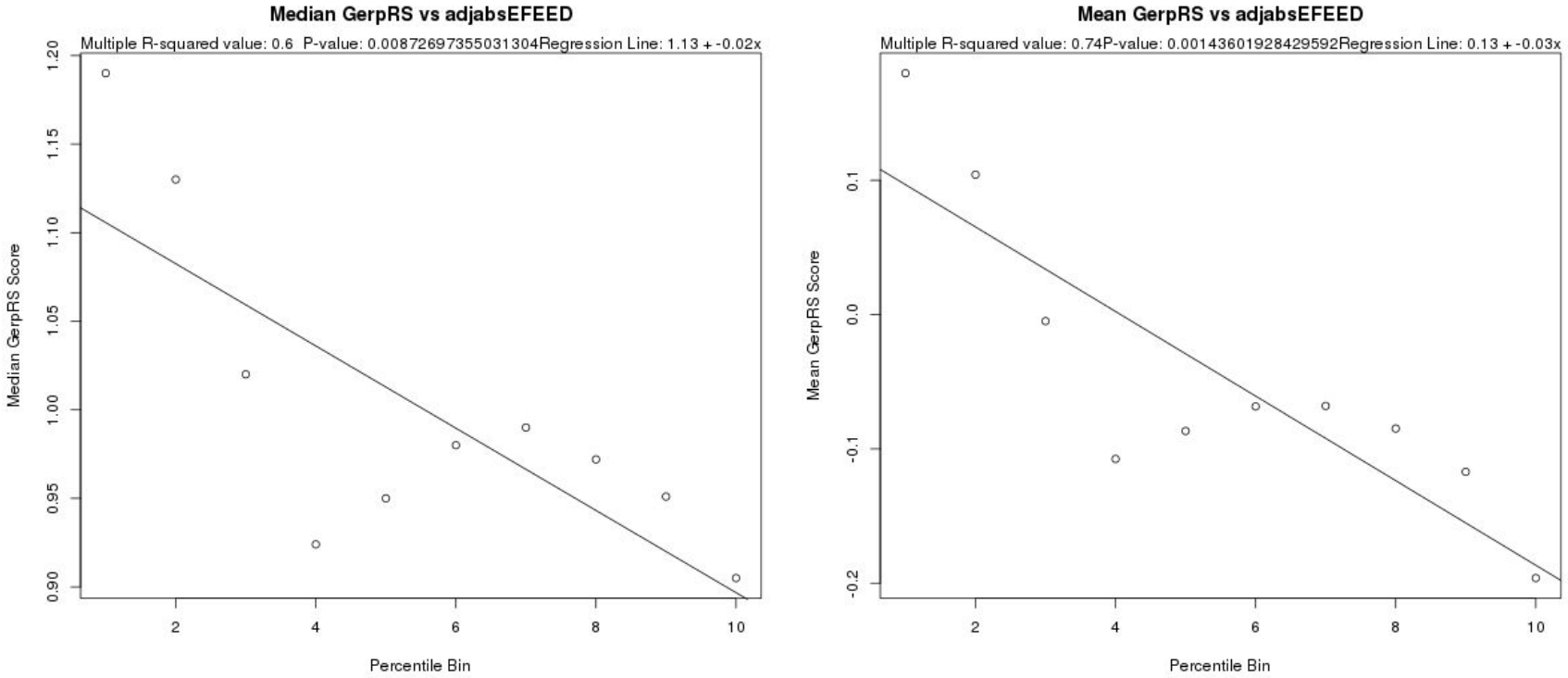
Supplementary Figure 72: % Non-zero Allele Frequency vs. Binned Change in Distance of the Ensemble of Structures to the Centroid (dCD) for Synonymous Variants



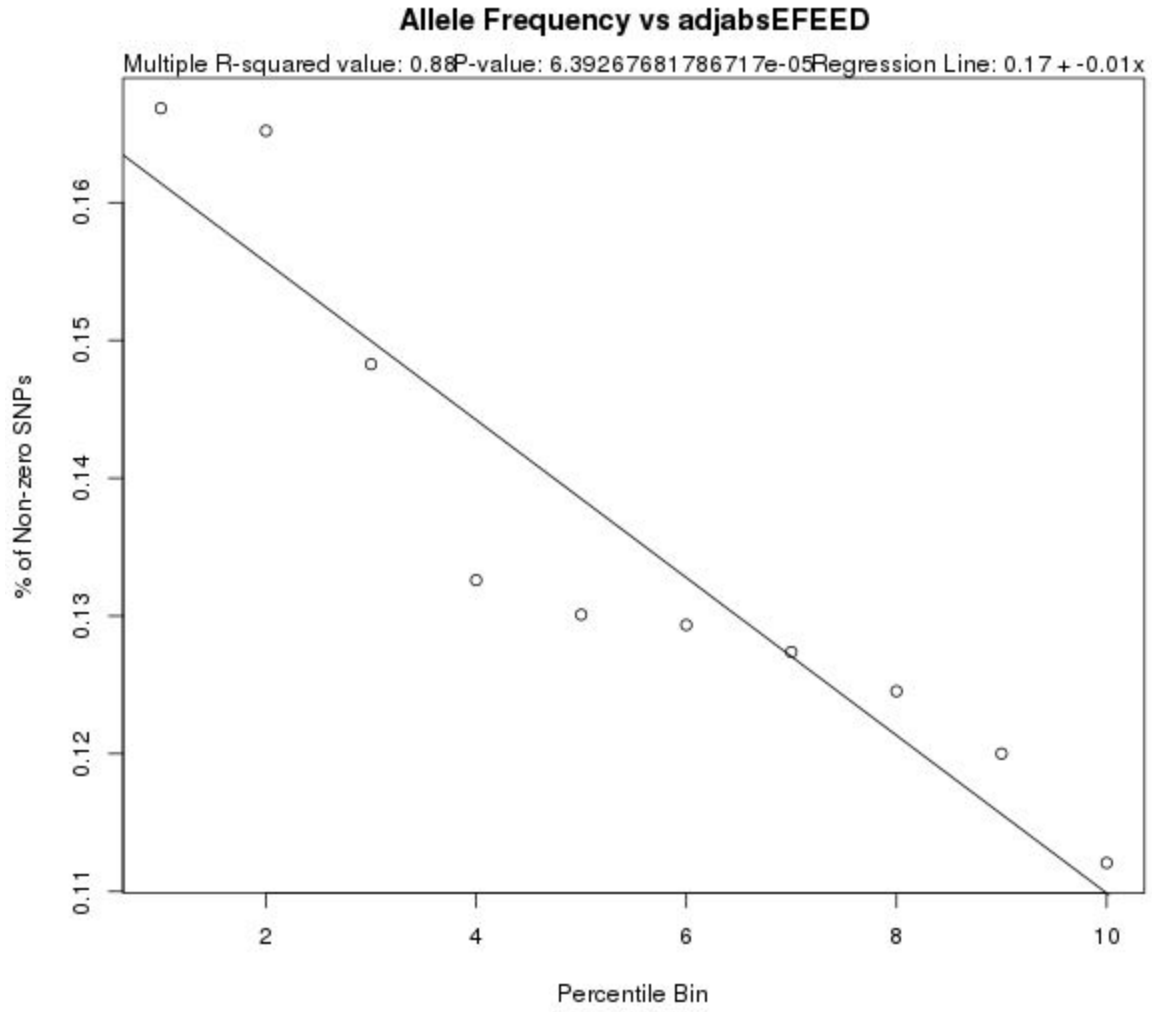
Supplementary Figure 73: Mean/Median GERP Score vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for Synonymous Variants



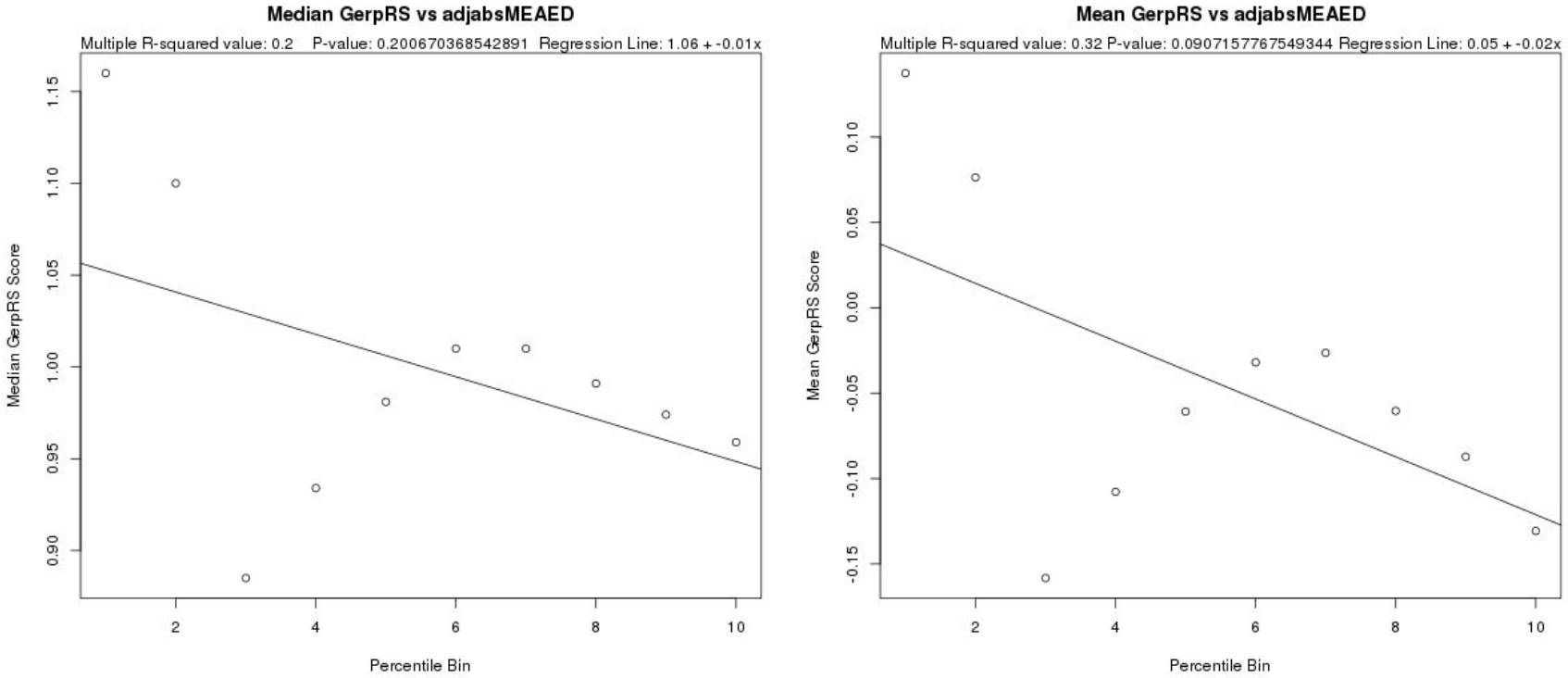
Supplementary Figure 74: % Non-zero Allele Frequency vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for Synonymous Variants



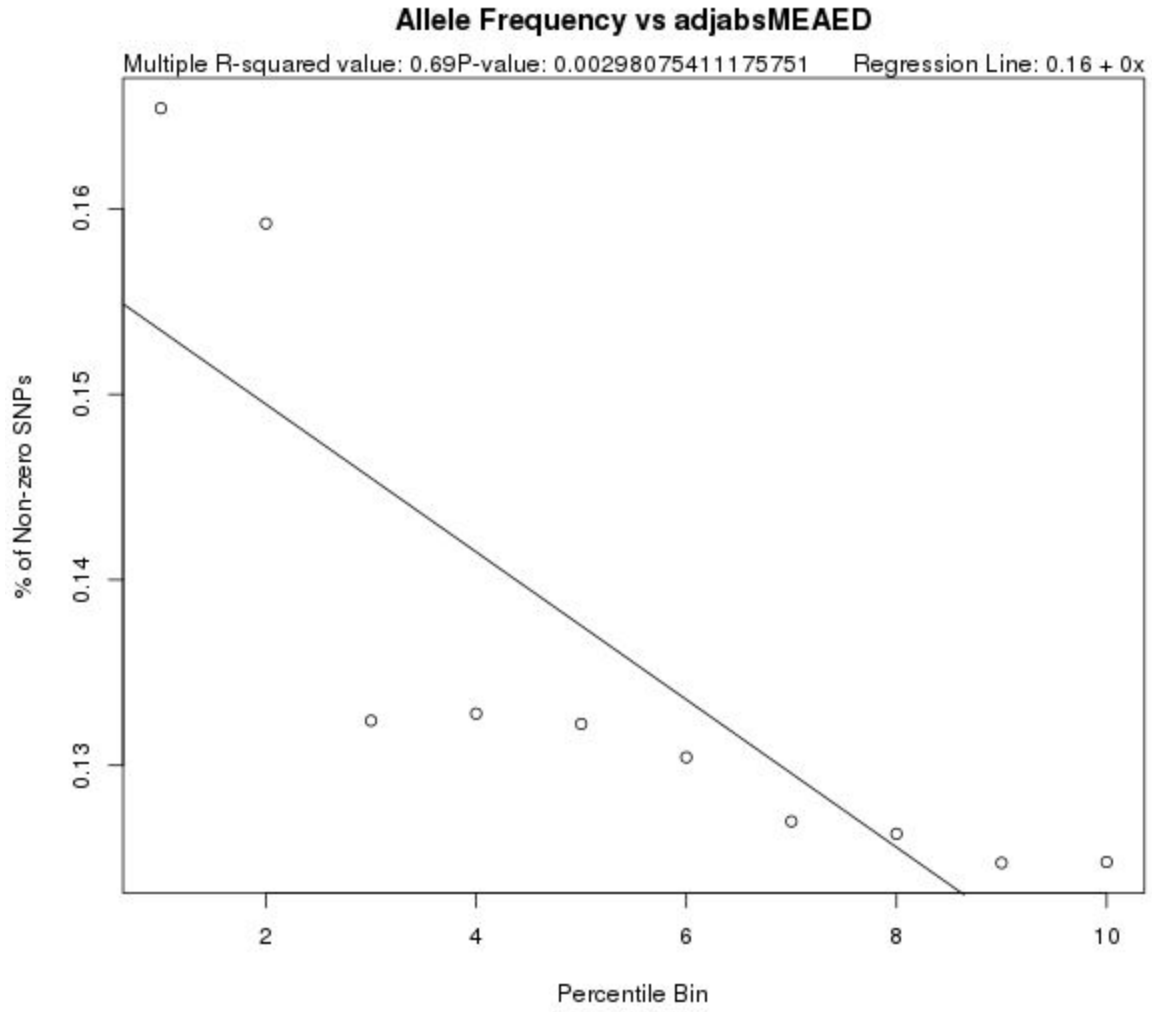
Supplementary Figure 75: Mean/Median GERP Score vs. Edit Distance Between Ensembles (EFEED) for Synonymous Variants



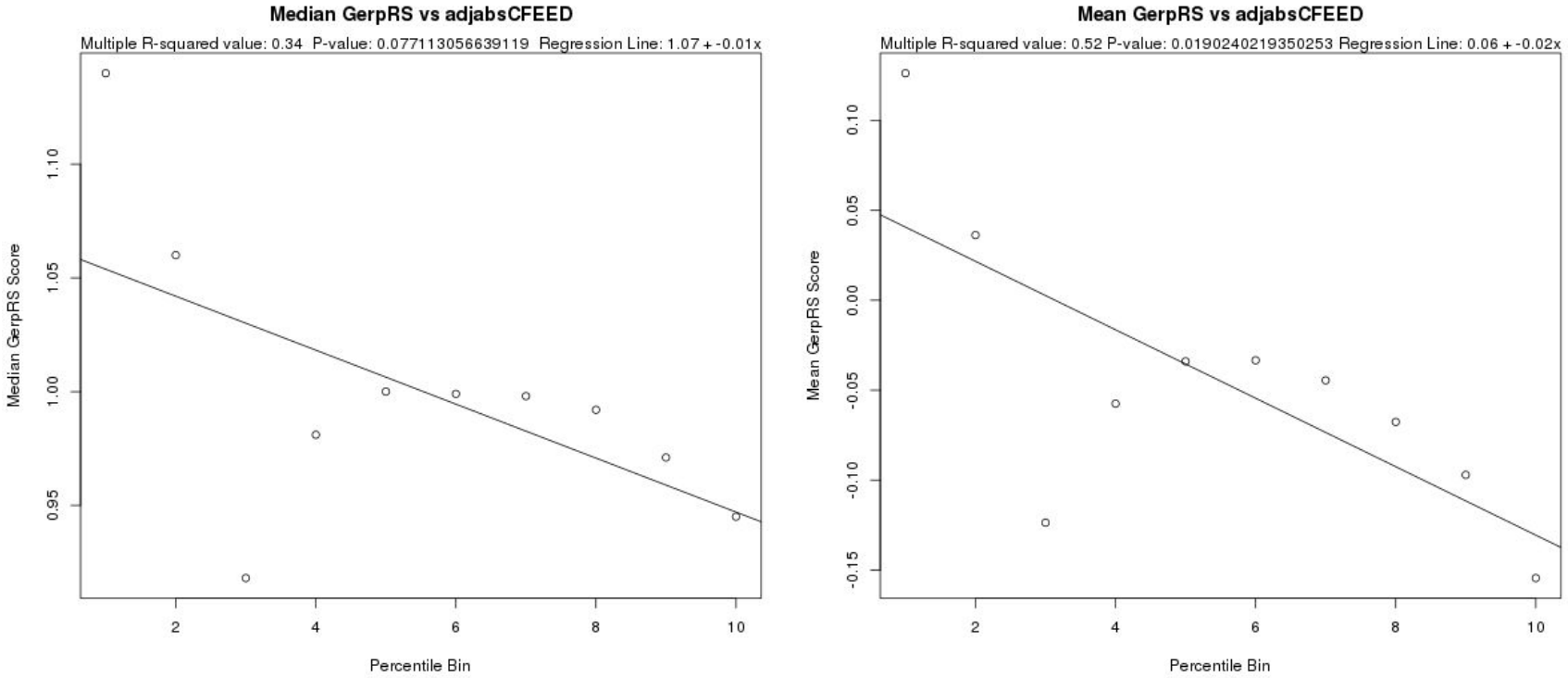
Supplementary Figure 76: % Non-zero Allele Frequency vs. Edit Distance Between Ensembles (EFEED) for Synonymous Variants



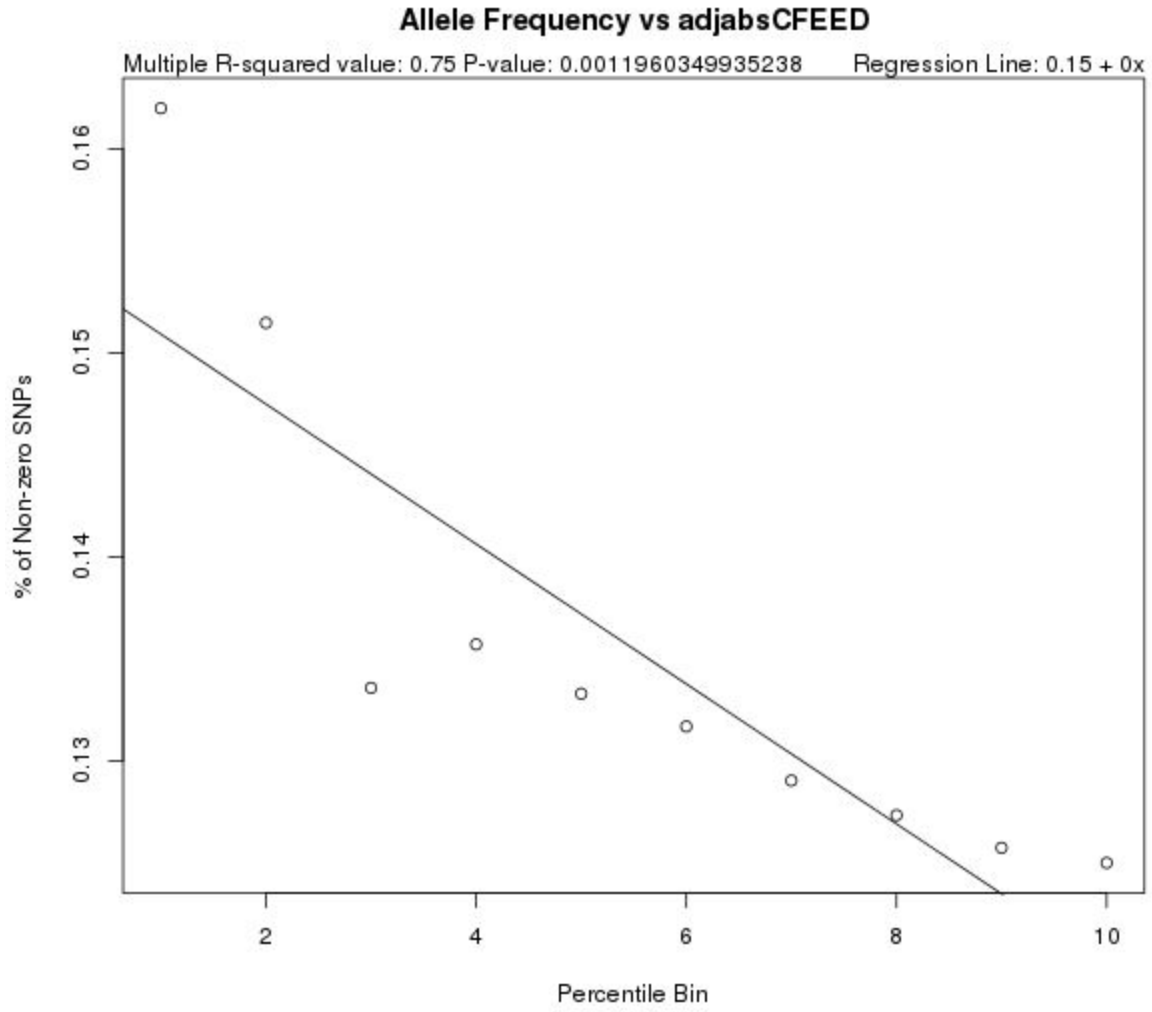
Supplementary Figure 77: Mean/Median GERP Score vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for Synonymous Variants



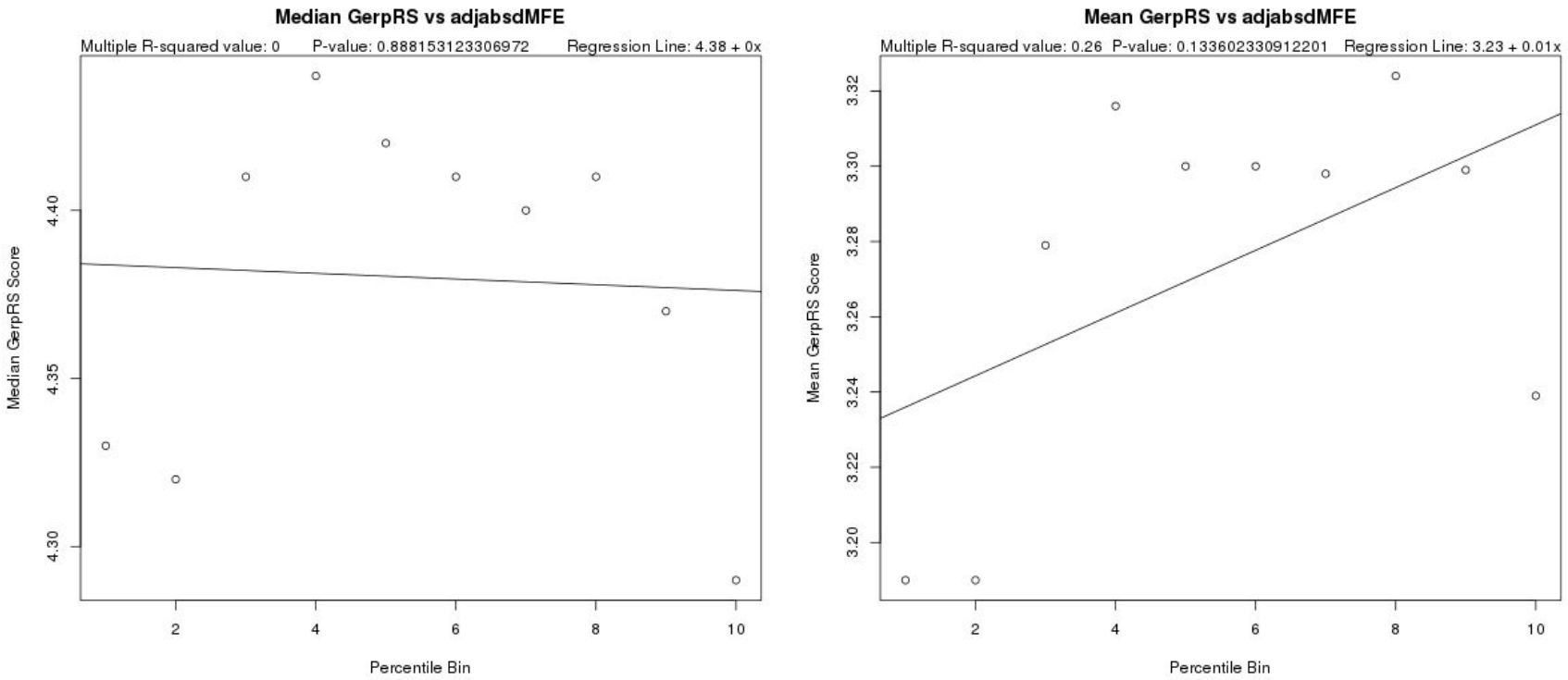
Supplementary Figure 78: % Non-zero Allele Frequency vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for Synonymous Variants



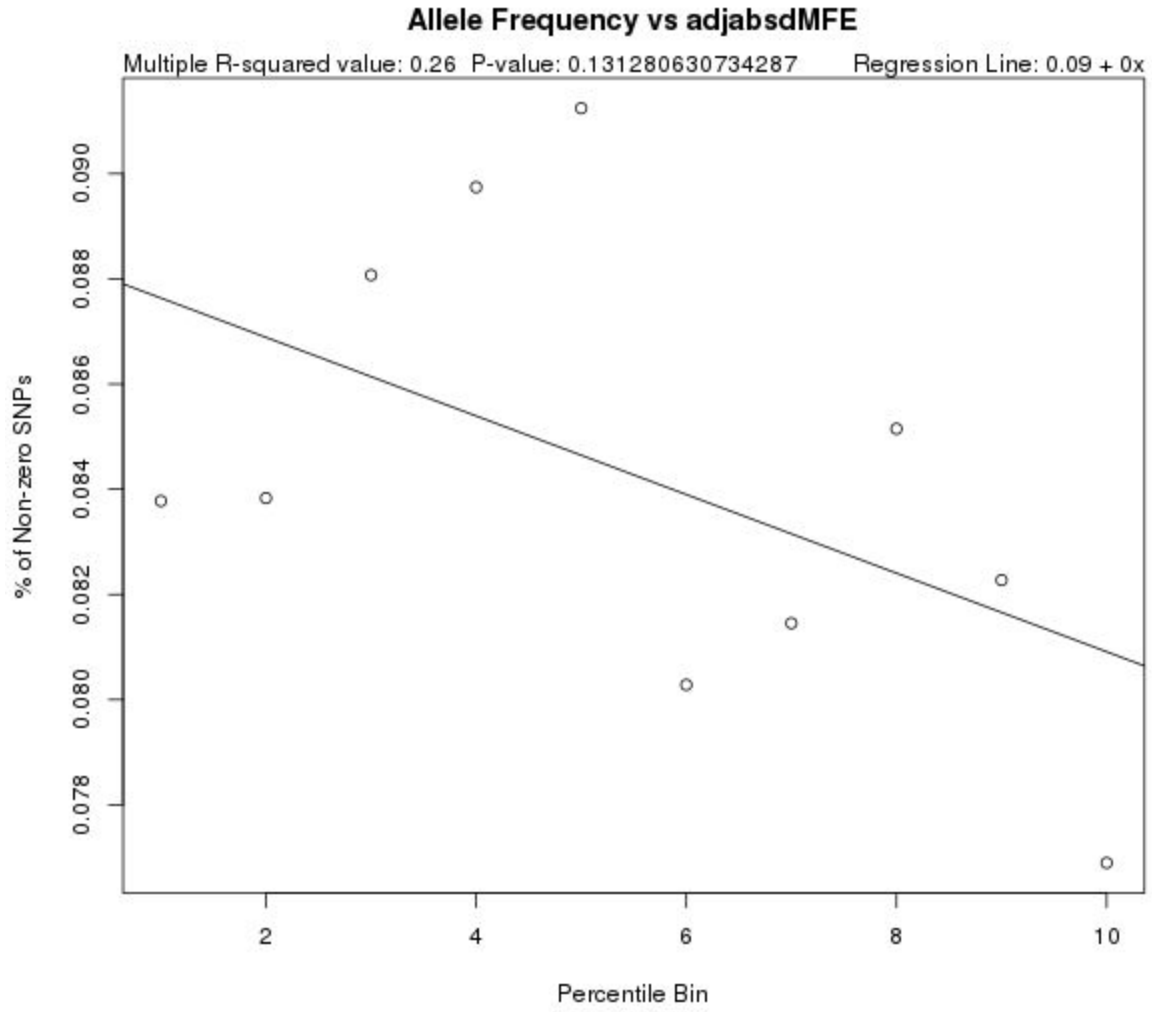
Supplementary Figure 79: Mean/Median GERP Score vs. Edit Distance Between Centroid Structures (CFEED) for Synonymous Variants



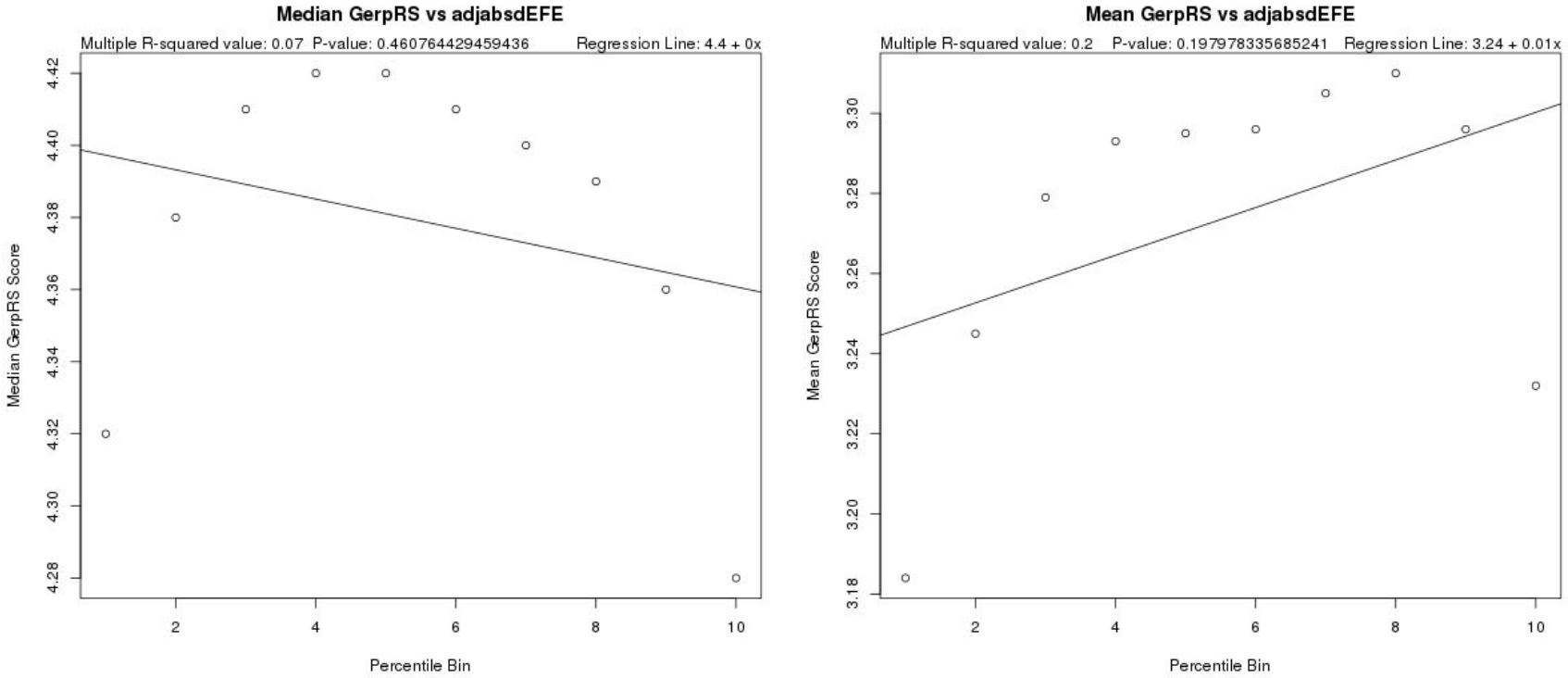
Supplementary Figure 80: % Non-zero Allele Frequency vs. Edit Distance Between Centroid Structures (CFEED) for Synonymous Variants



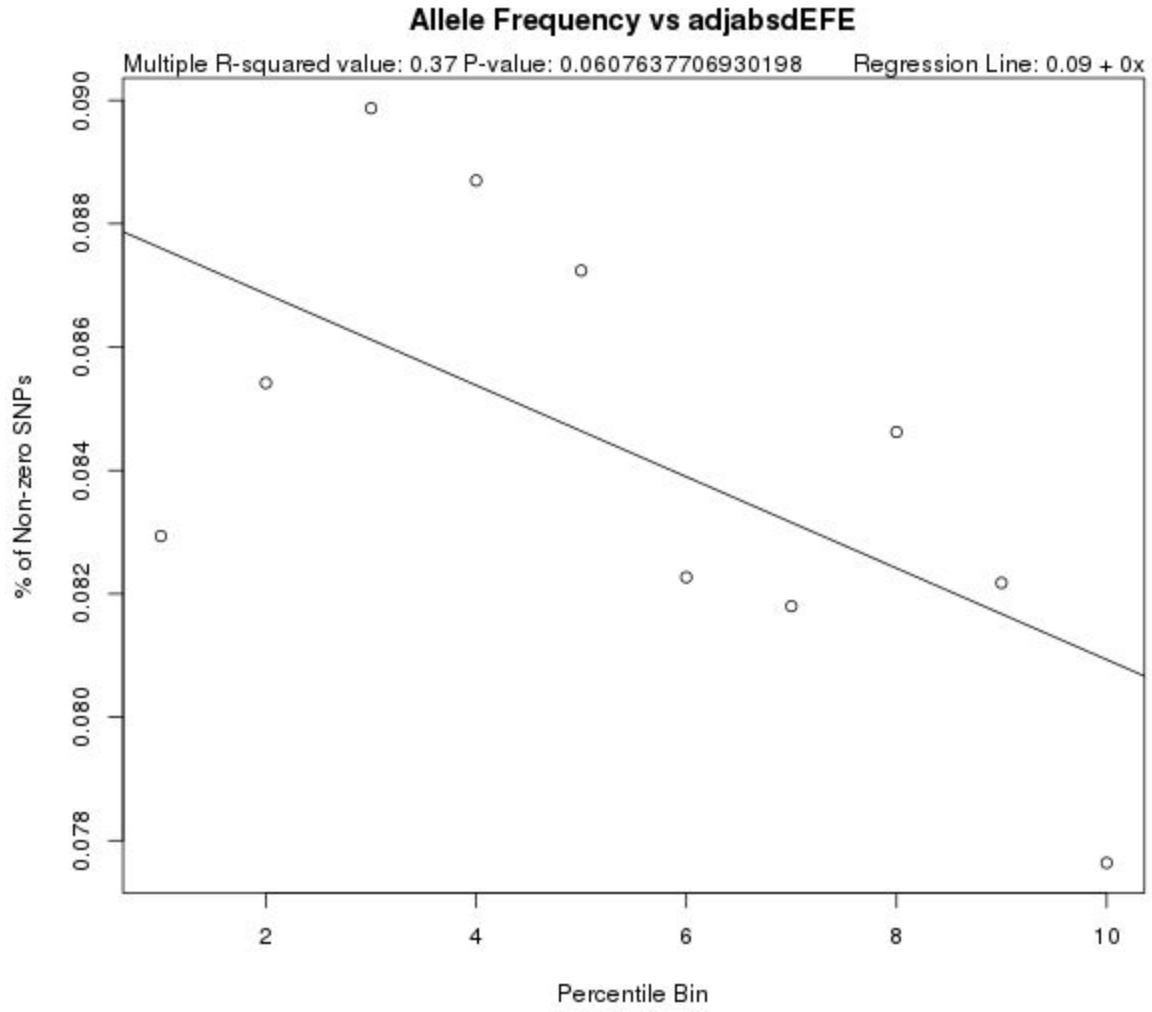
Supplementary Figure 81: Mean/Median GERP Score vs. Binned Change in Minimum Free Energy (dMFE) for Missense Variants



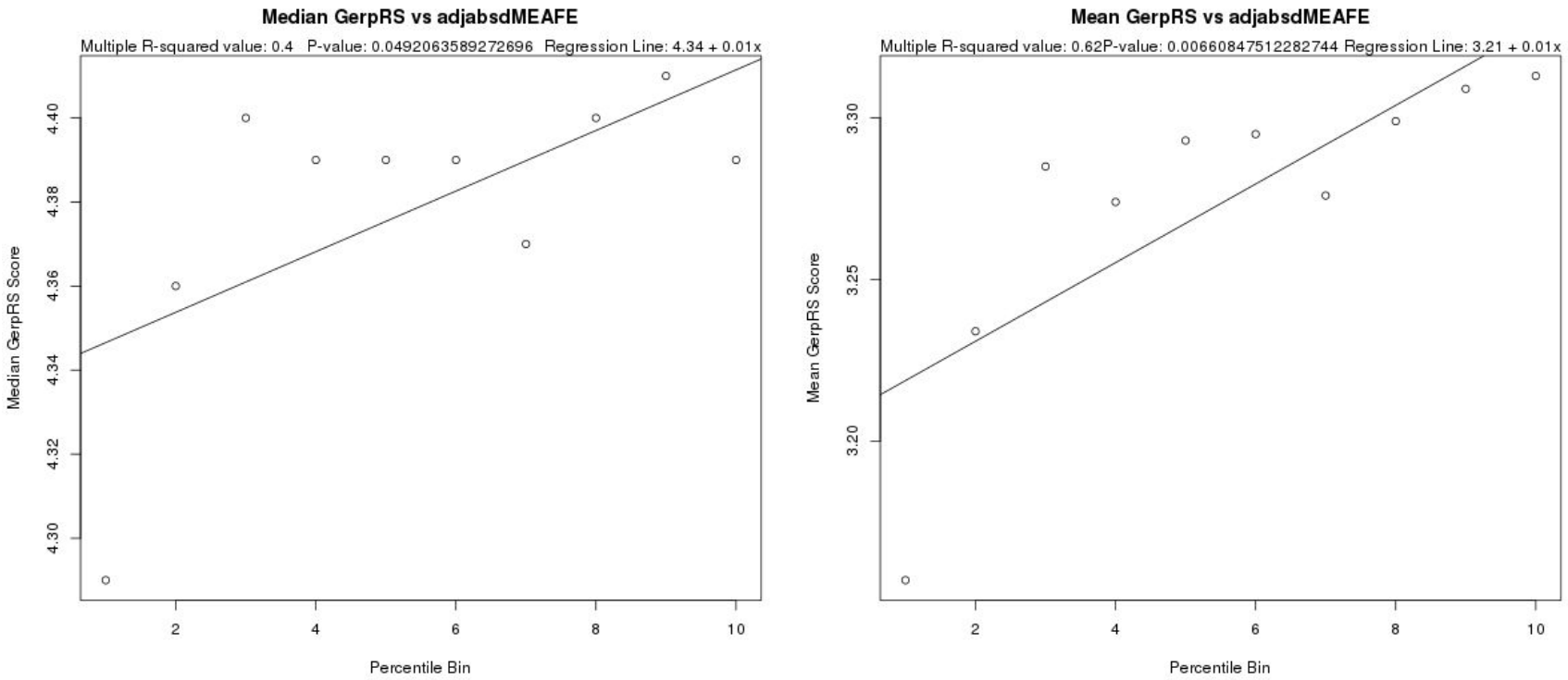
Supplementary Figure 82: % Non-zero Allele Frequency vs. Binned Change in Minimum Free Energy (dMFE) for Missense Variants



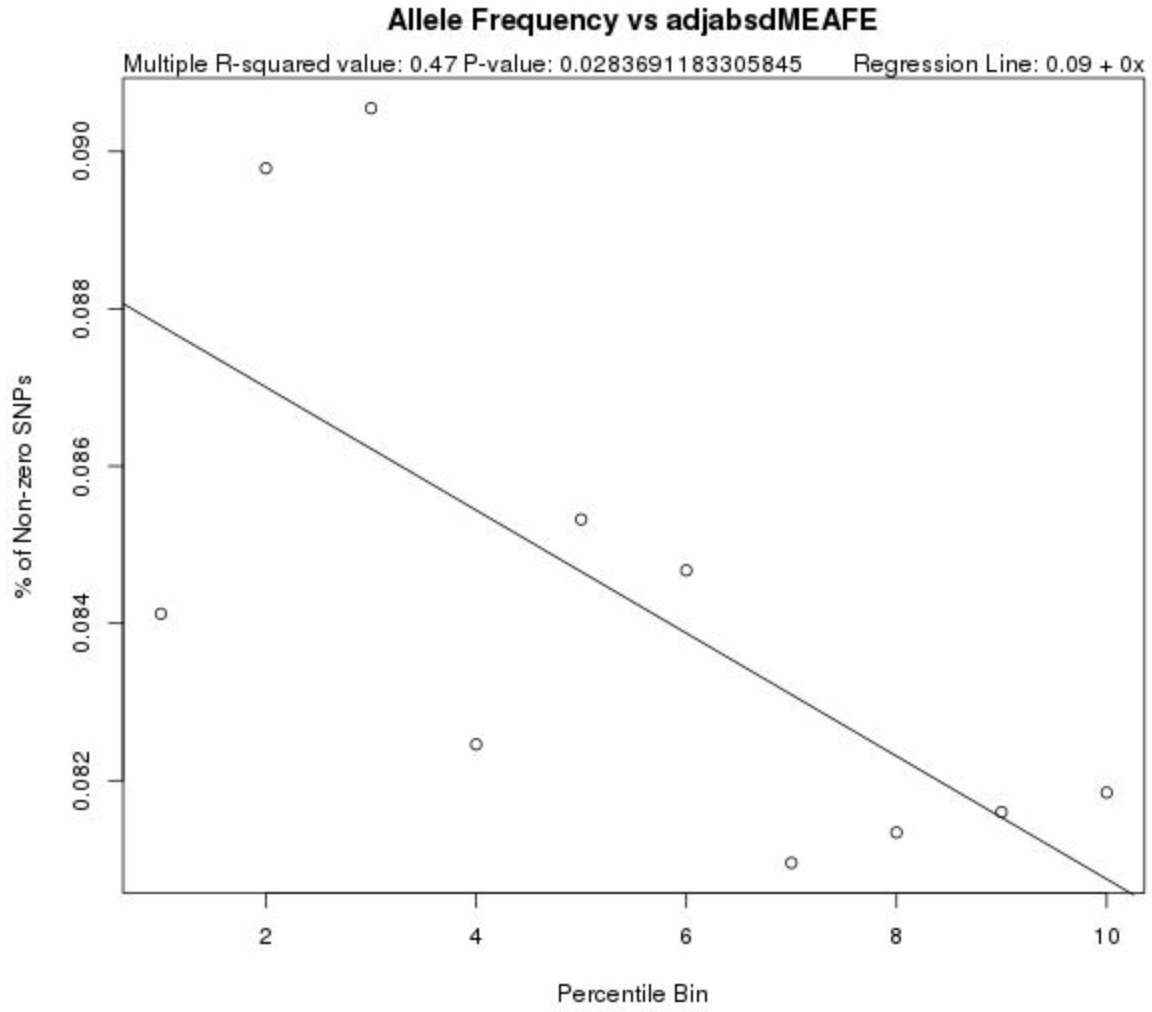
Supplementary Figure 83: Mean/Median GERP Score vs. Binned Change in Ensemble Free Energy (dEFE) for Missense Variants



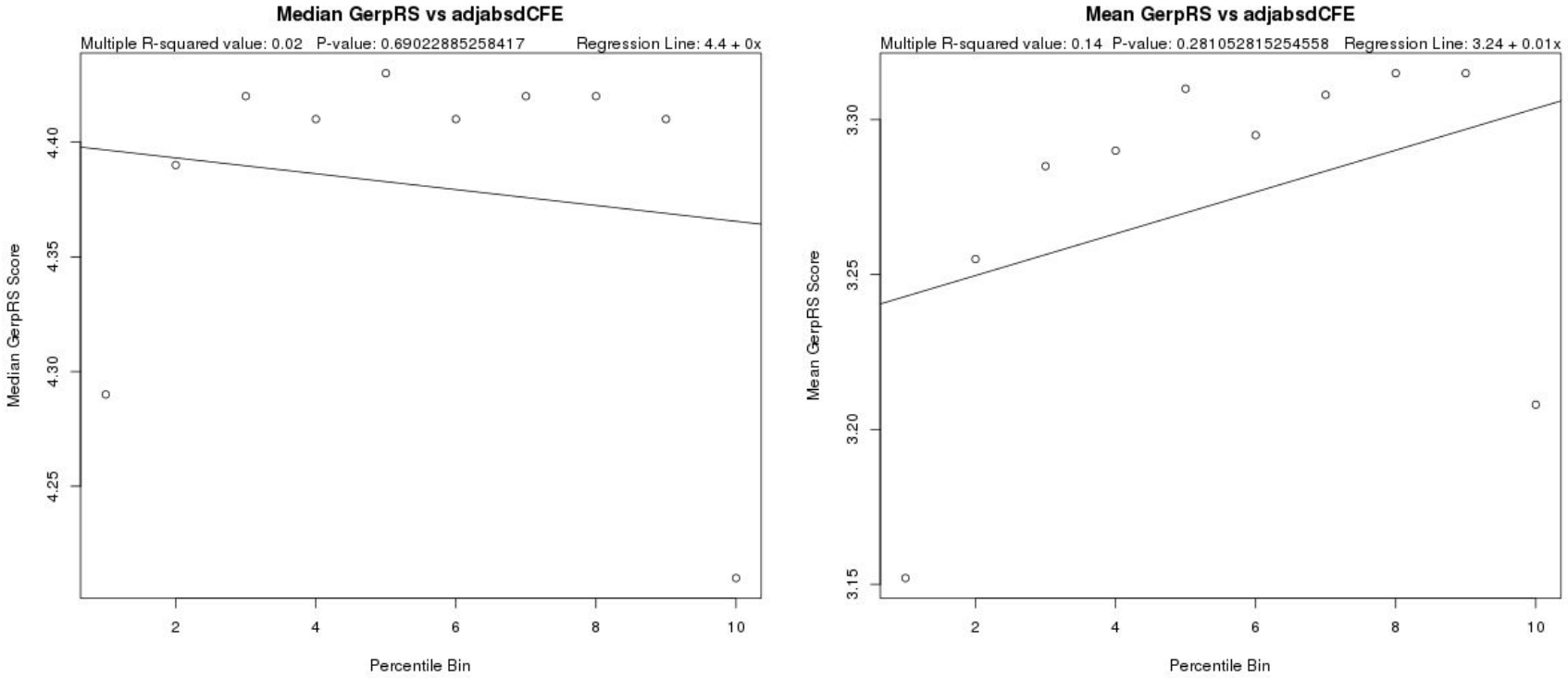
Supplementary Figure 84: % Non-zero Allele Frequency vs. Binned Change in Ensemble Free Energy (dEFE) for Missense Variants



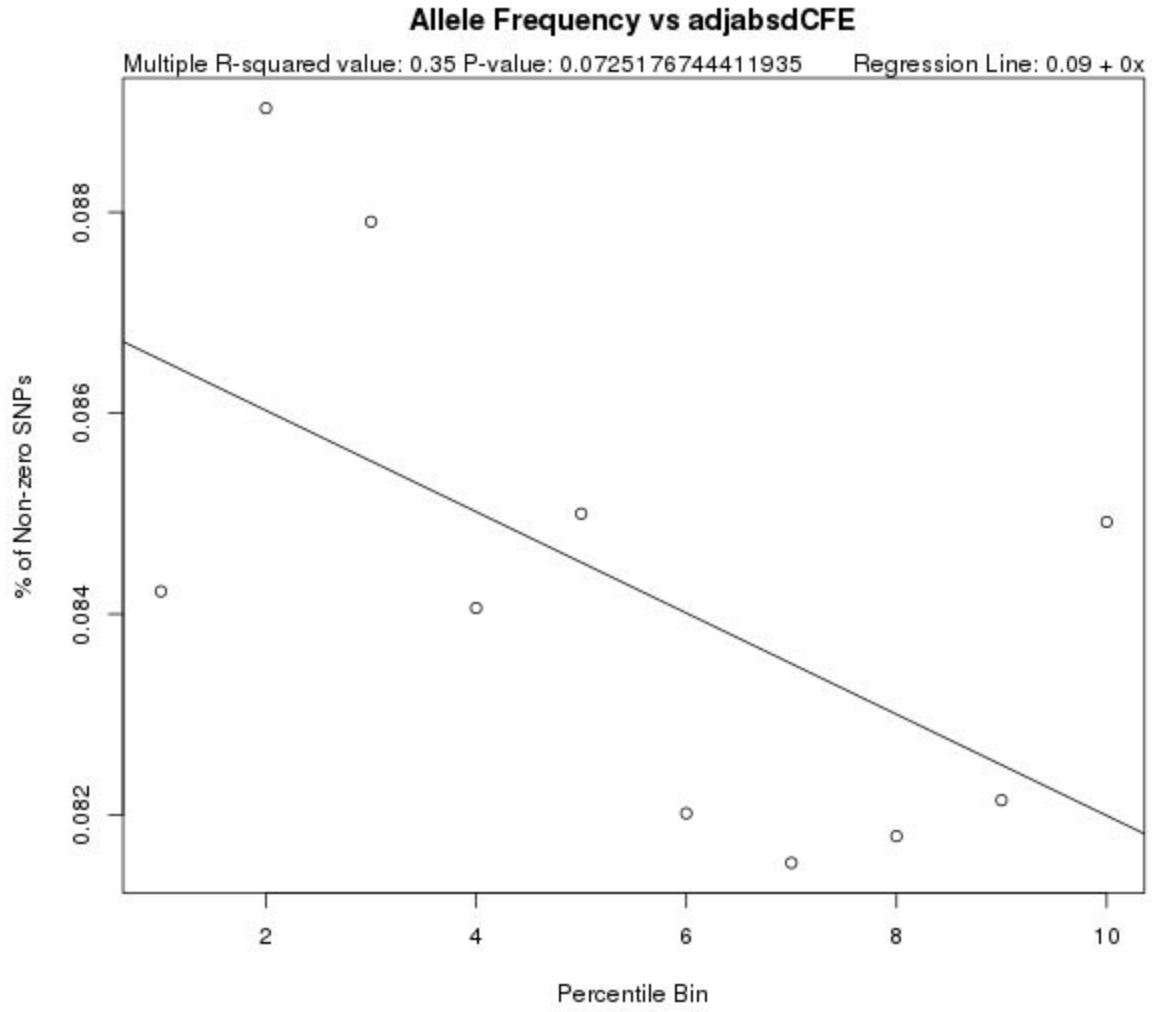
Supplementary Figure 85: Mean/Median GERP Score vs. Binned Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for Missense Variants



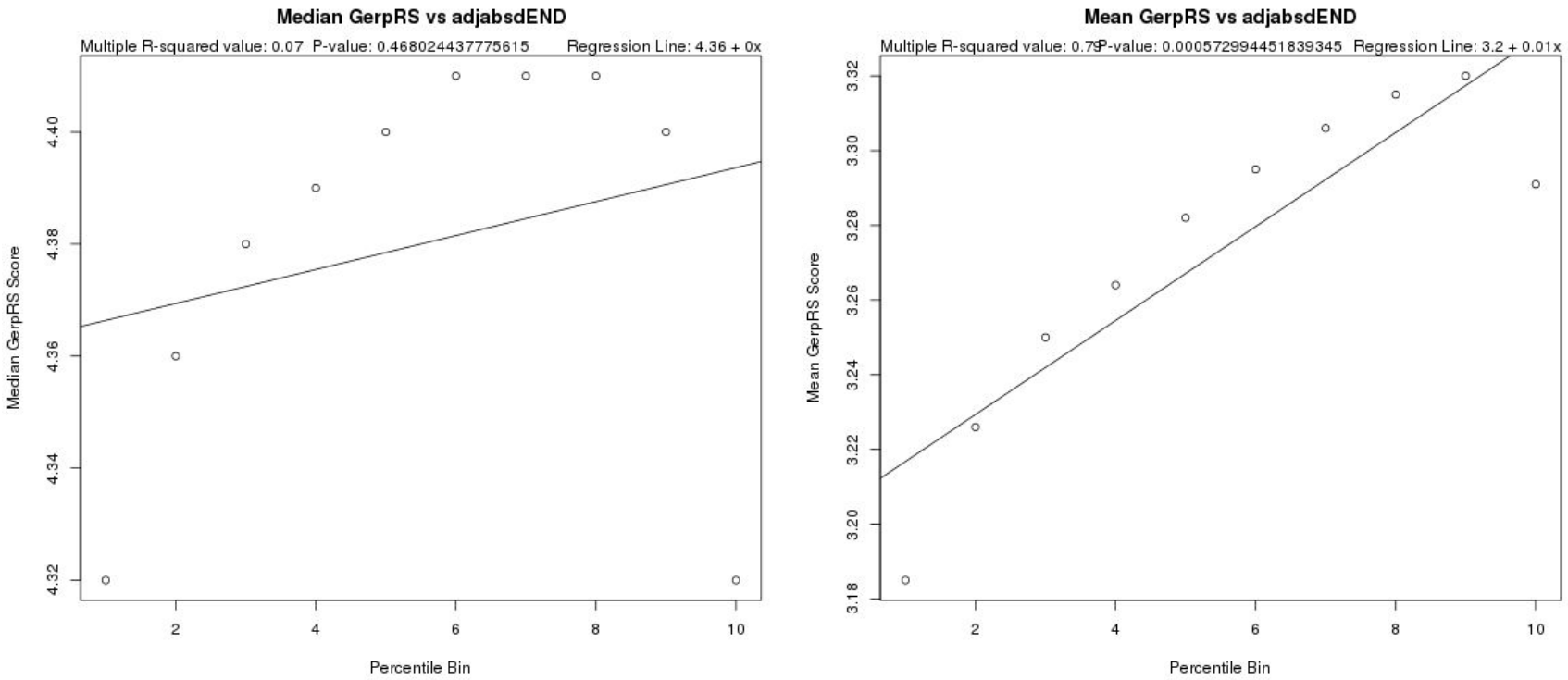
Supplementary Figure 86: % Non-zero Allele Frequency vs. Binned Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for Missense Variants



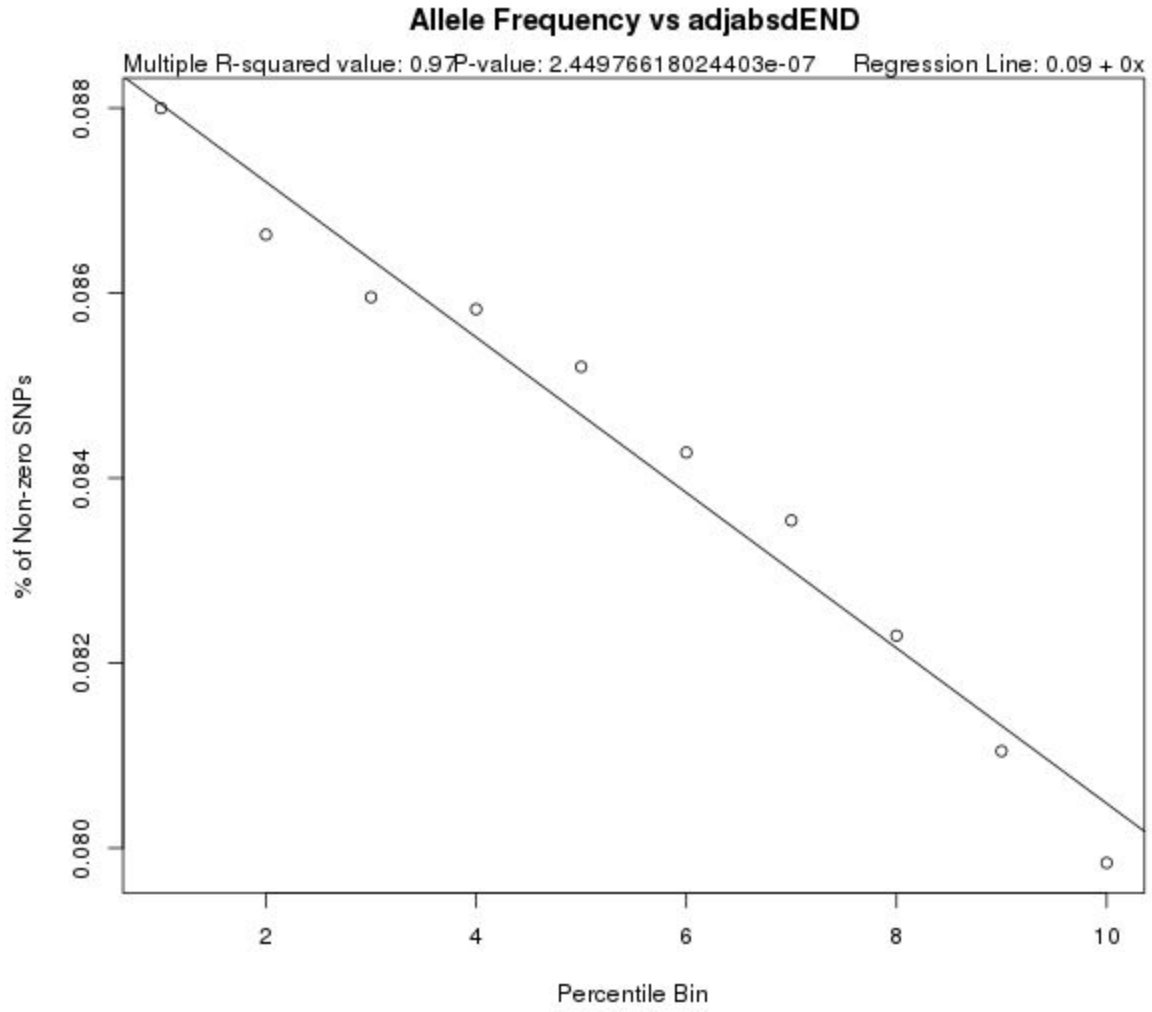
Supplementary Figure 87: Mean/Median GERP Score vs. Binned Change in Free Energy of the Centroid (dCFE) for Missense Variants



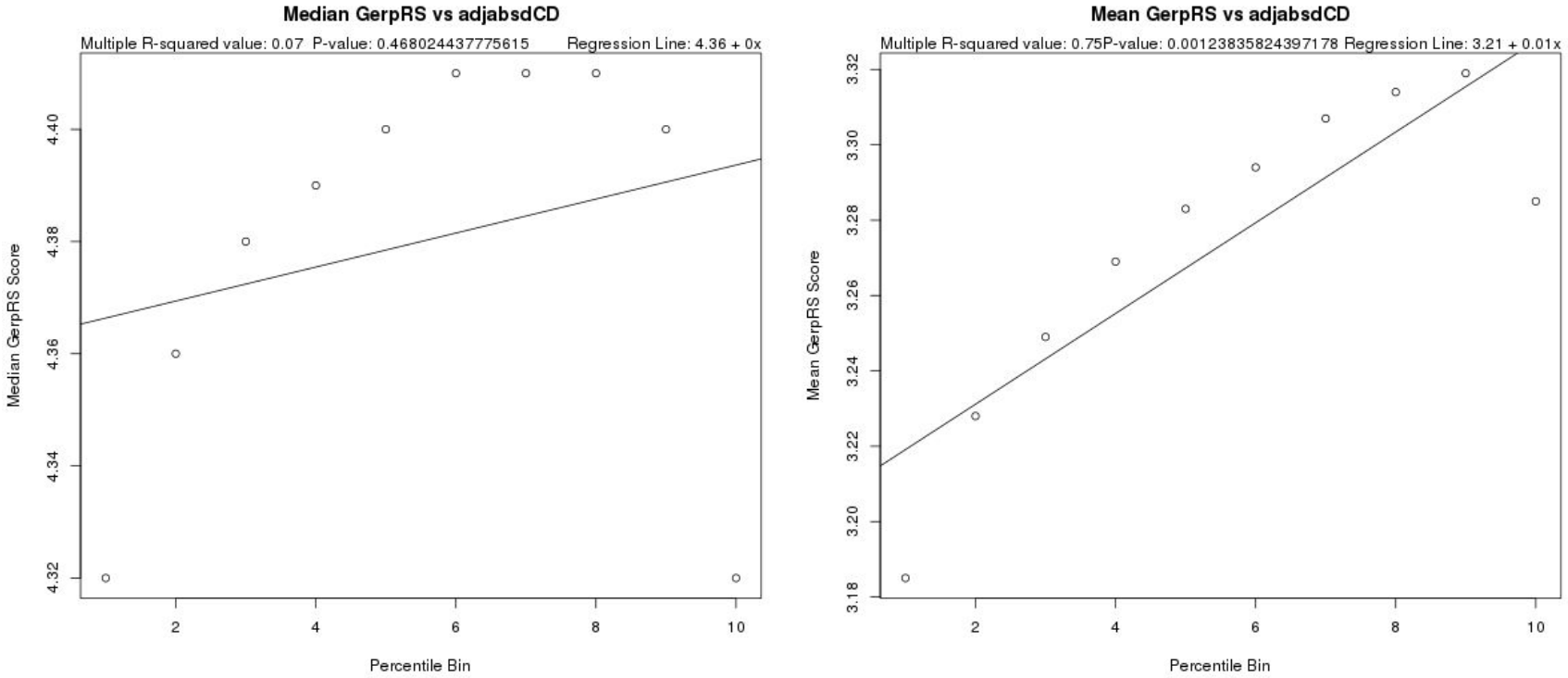
Supplementary Figure 88: % Non-zero Allele Frequency vs. Binned Change in Free Energy of the Centroid (dCFE) for Missense Variants



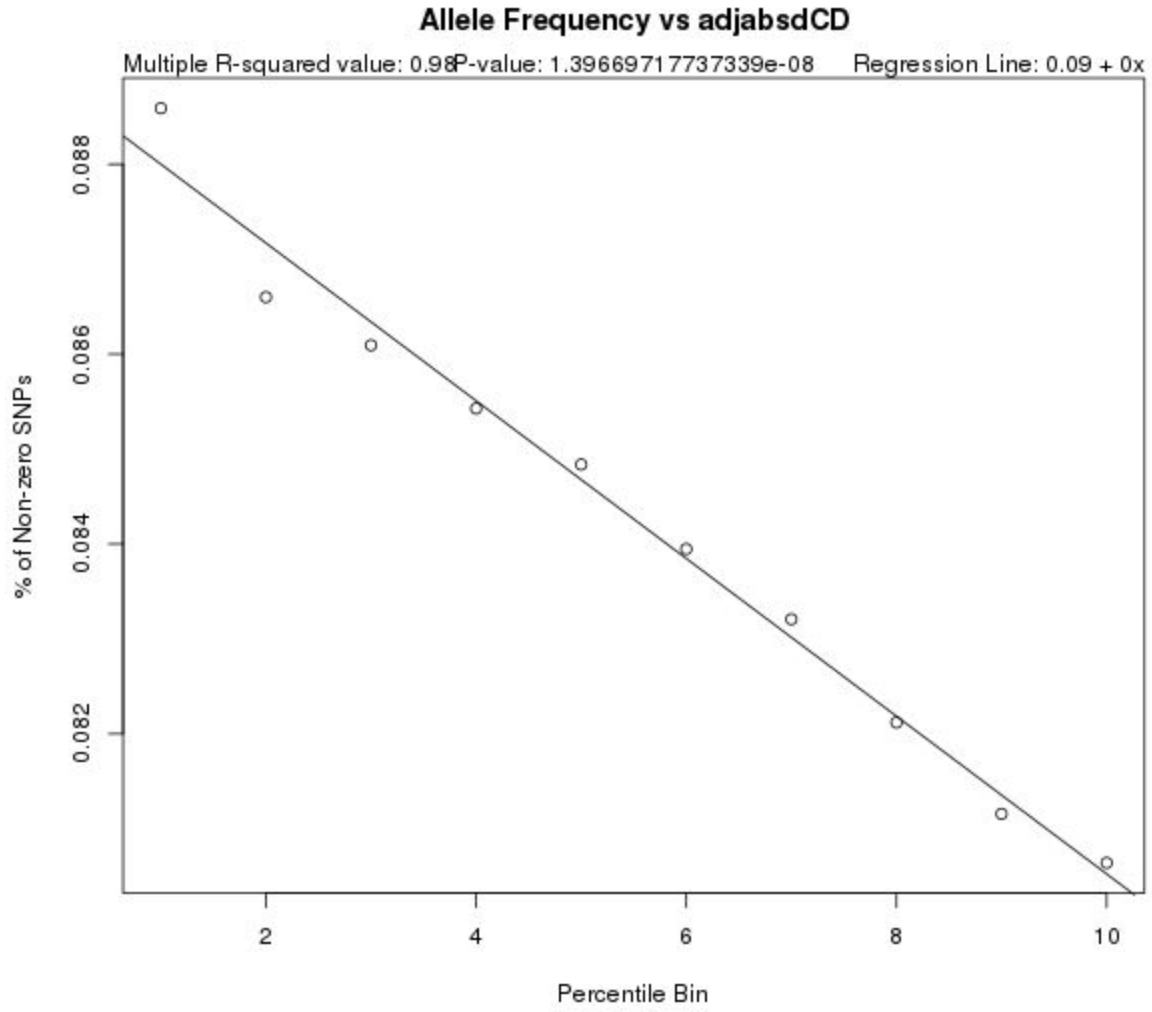
Supplementary Figure 89: Mean/Median GERP Score vs. Binned Change in Ensemble Diversity (dEND) for Missense Variants



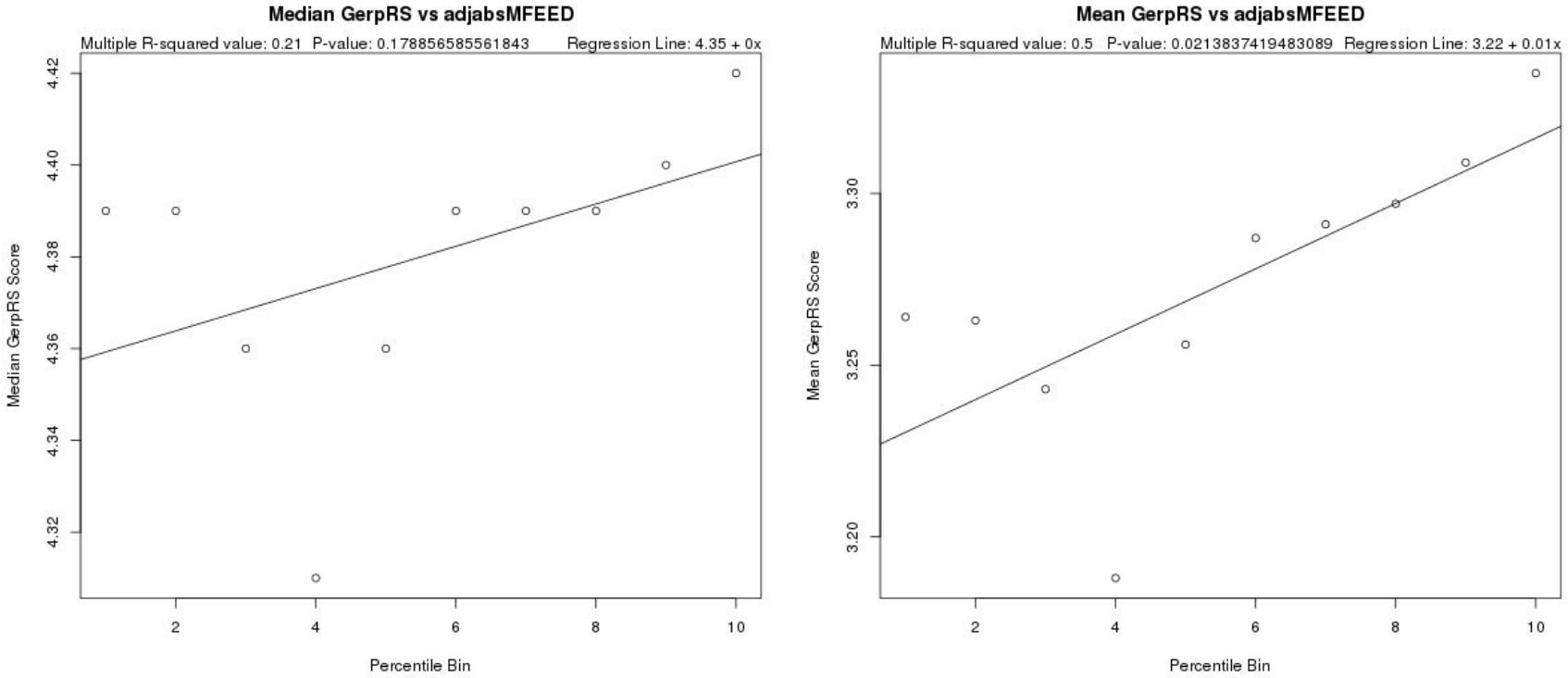
Supplementary Figure 90: % Non-zero Allele Frequency vs. Binned Change in Ensemble Diversity (dEND) for Missense Variants



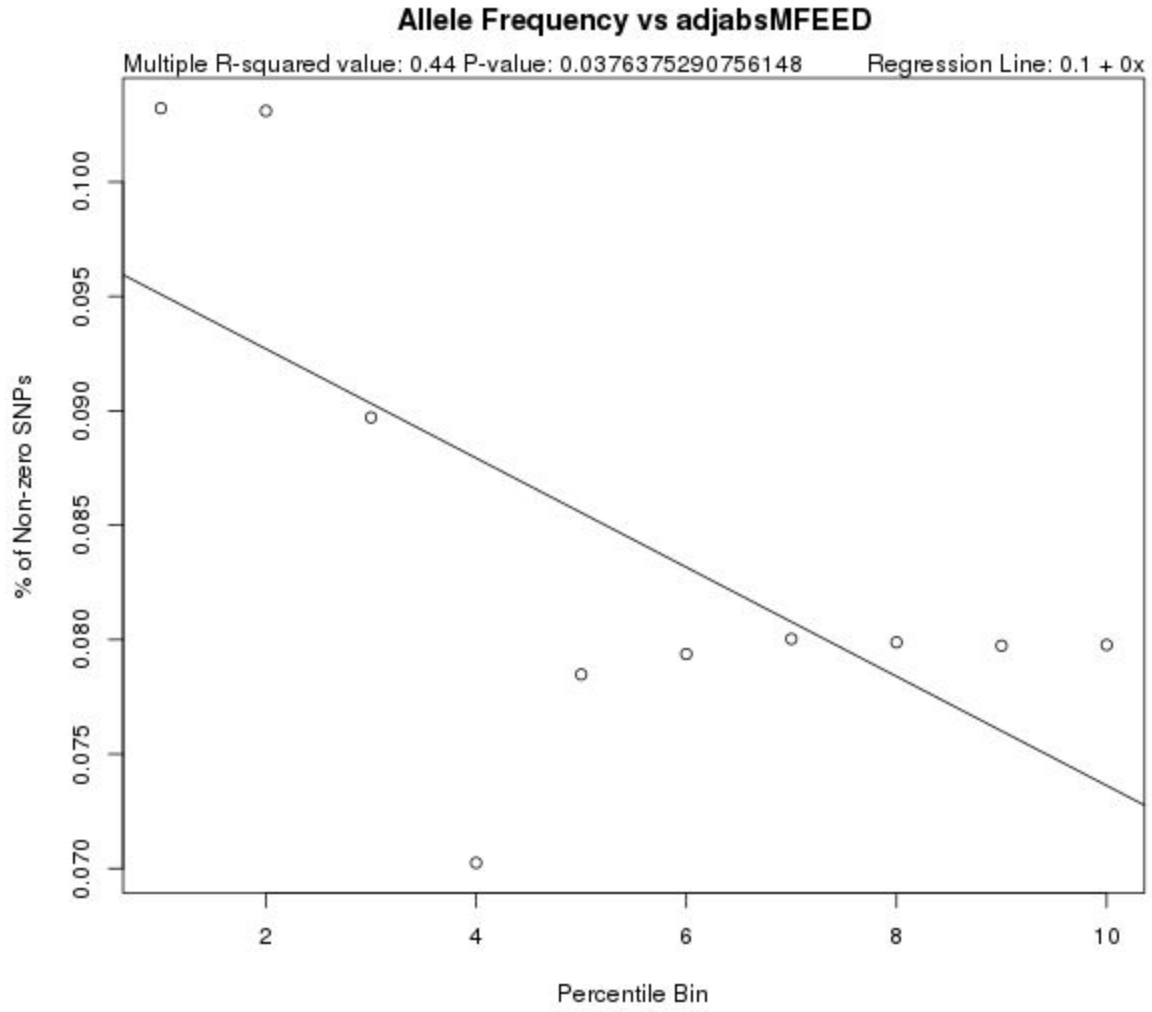
Supplementary Figure 91: Mean/Median GERP Score vs. Binned Change in Distance of the Ensemble of Structures to the Centroid (dCD) for Missense Variants



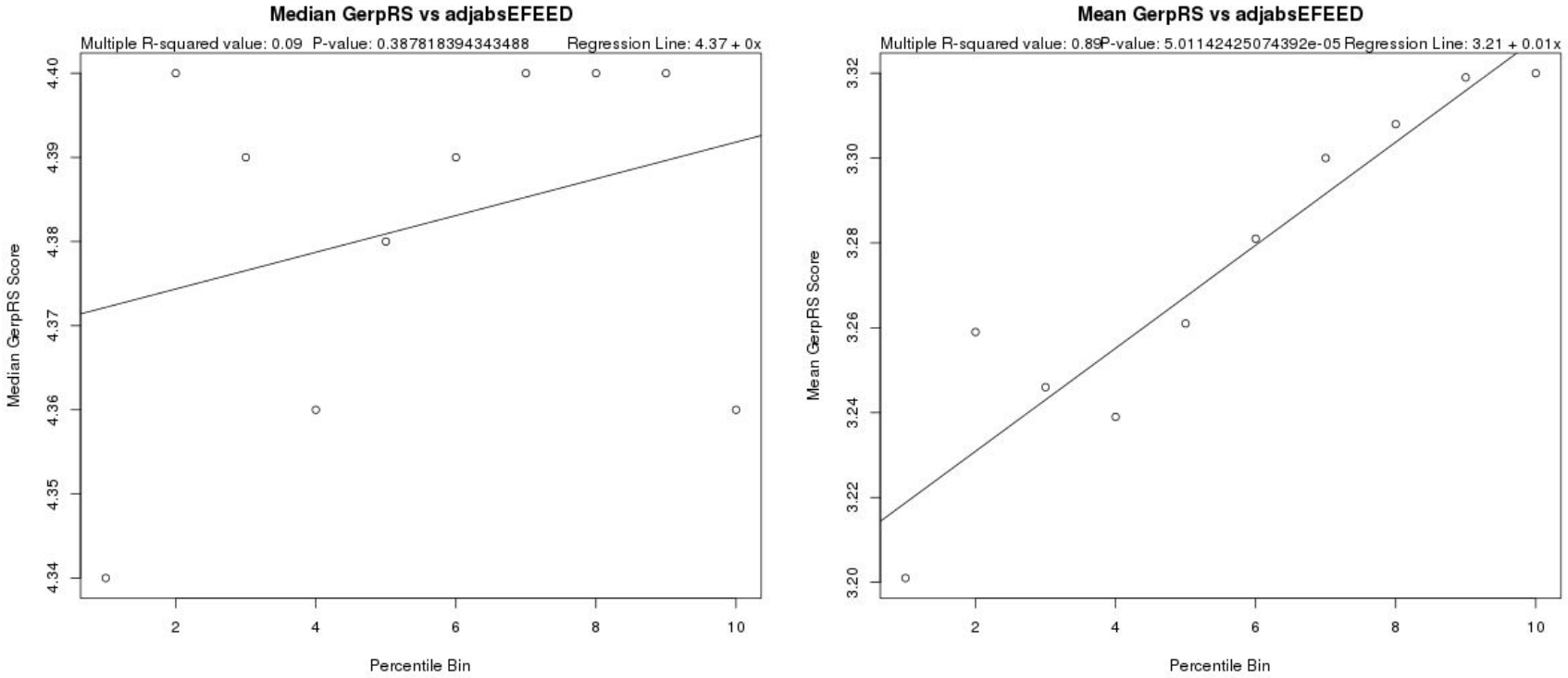
Supplementary Figure 92: % Non-zero Allele Frequency vs. Binned Change in Distance of the Ensemble of Structures to the Centroid (dCD) for Missense Variants



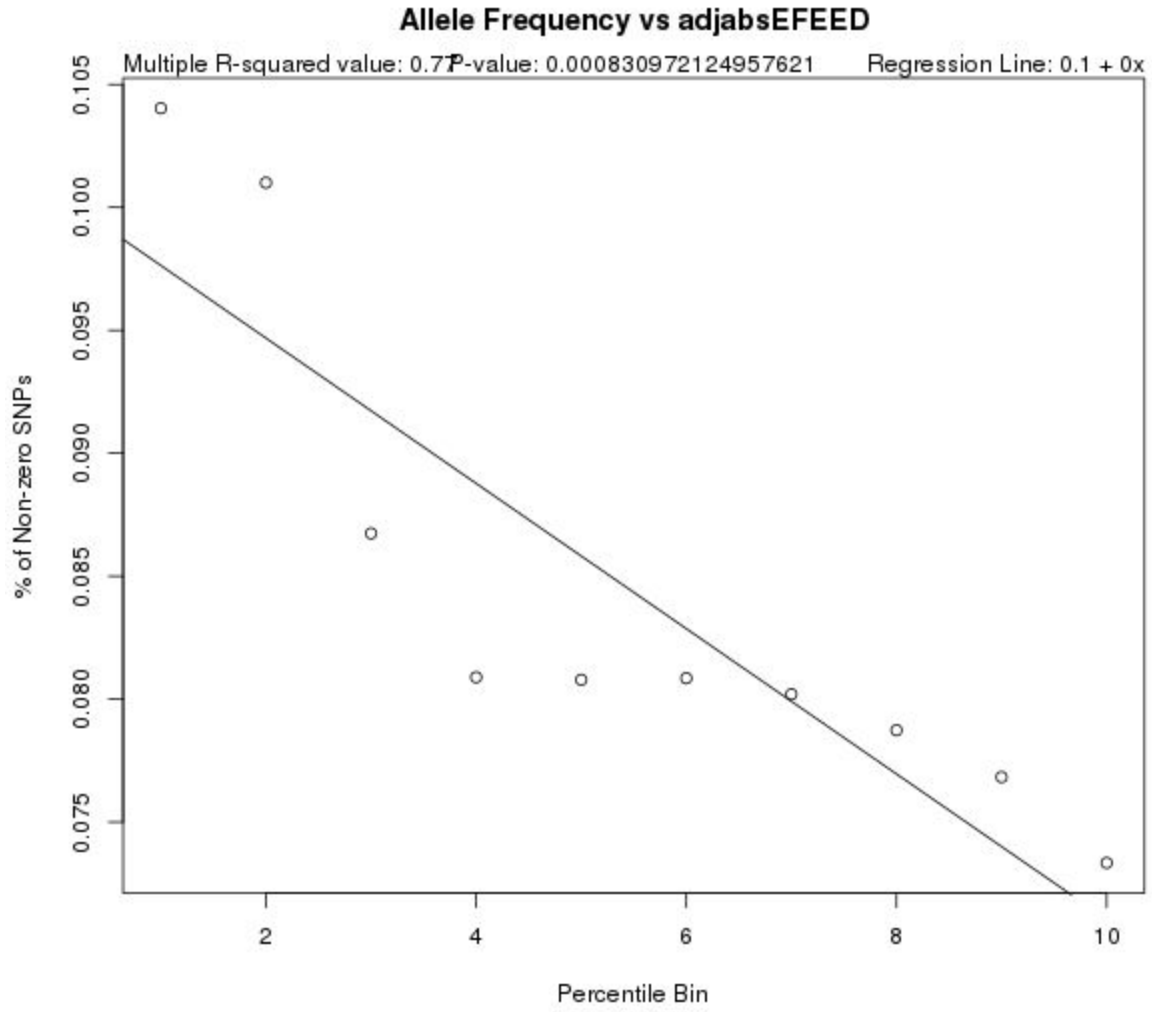
Supplementary Figure 93: Mean/Median GERP Score vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for Missense Variants



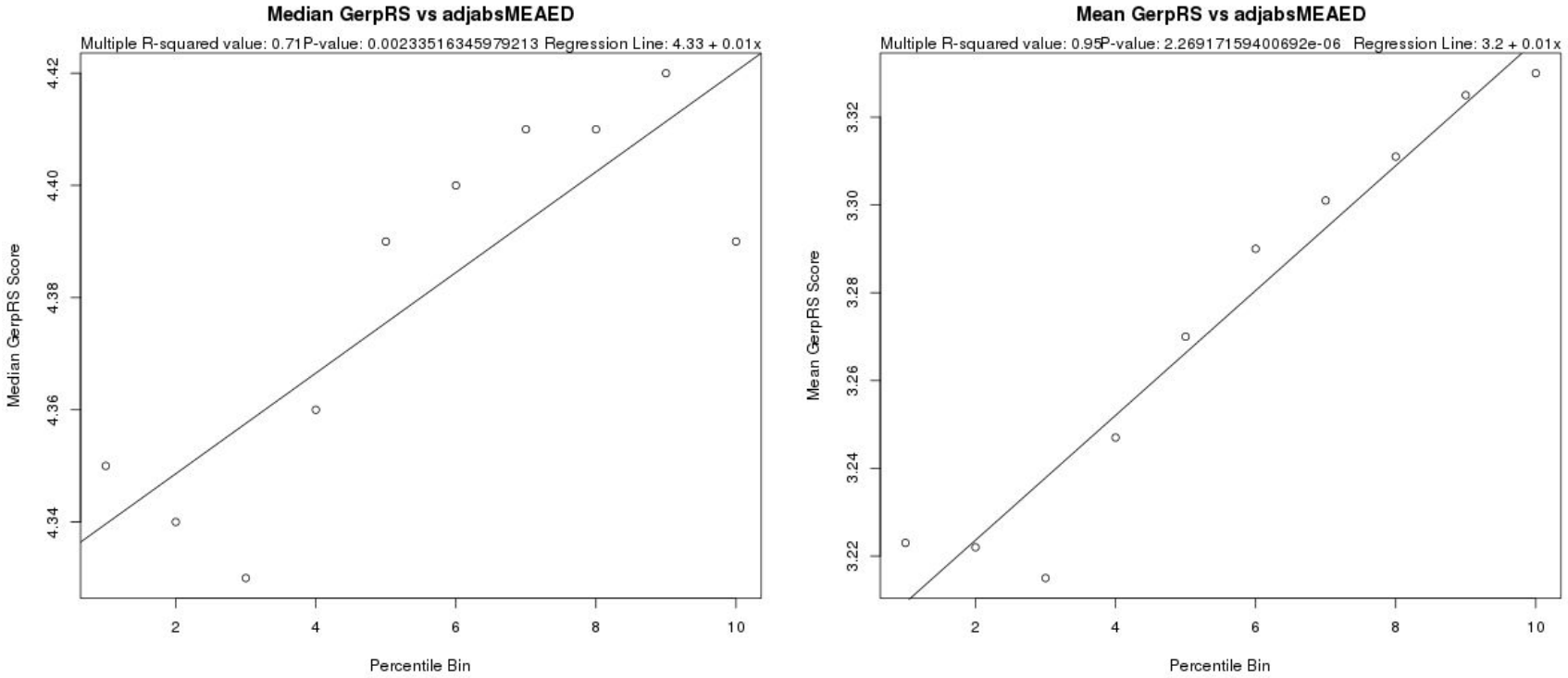
Supplementary Figure 94: % Non-zero Allele Frequency vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for Missense Variants



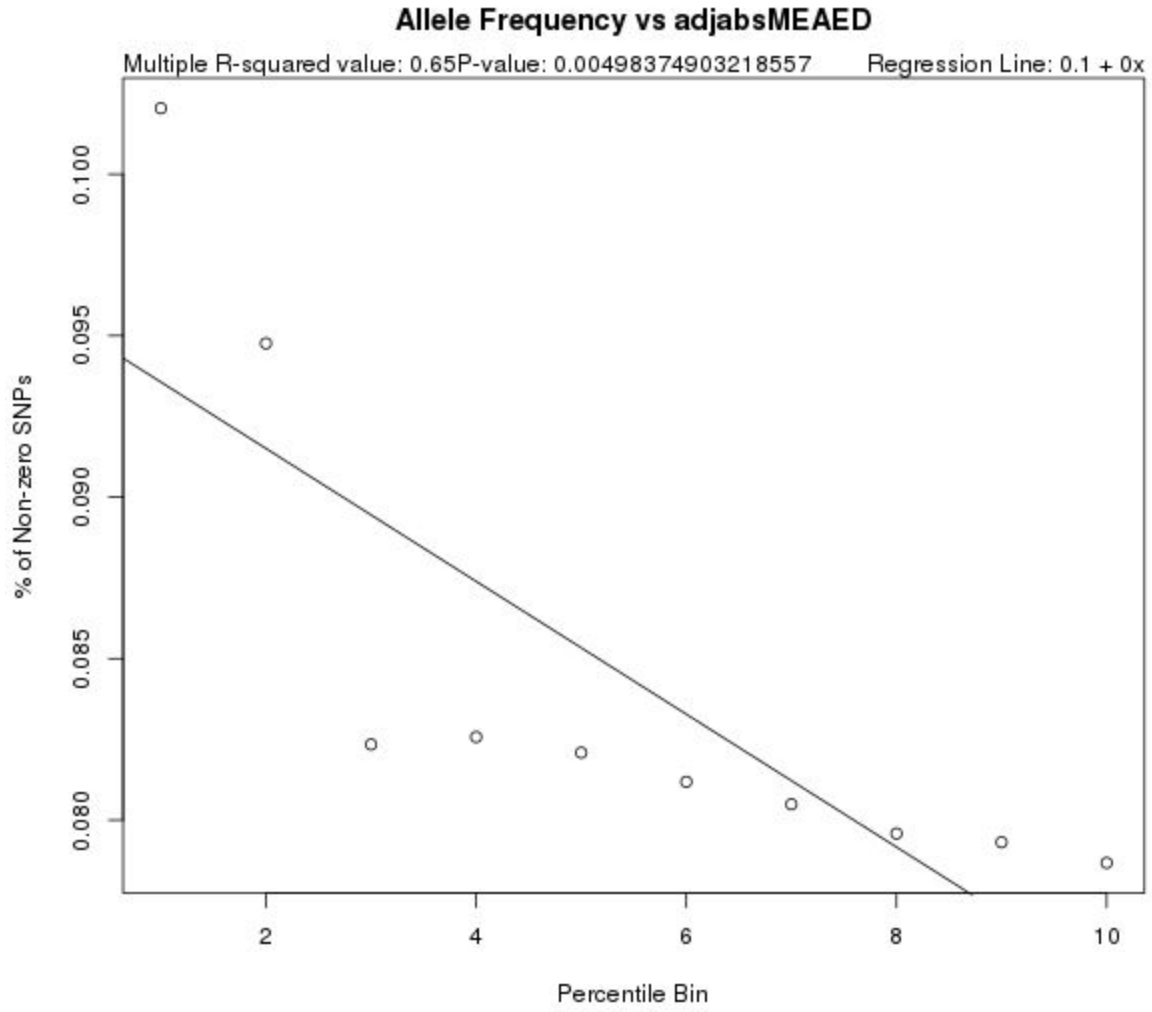
Supplementary Figure 95: Mean/Median GERP Score vs. Edit Distance Between Ensembles (EFEED) for Missense Variants



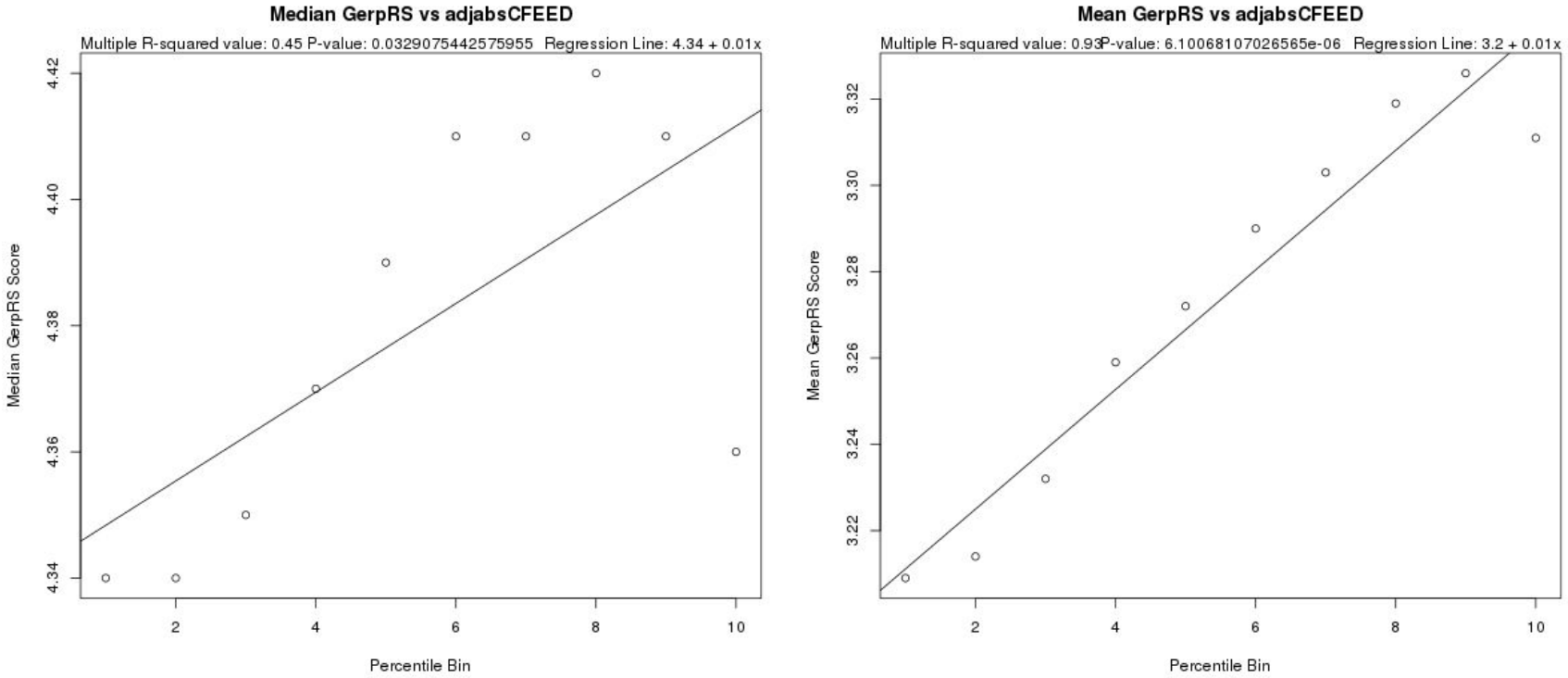
Supplementary Figure 96: % Non-zero Allele Frequency vs. Edit Distance Between Ensembles (EFEED) for Missense Variants



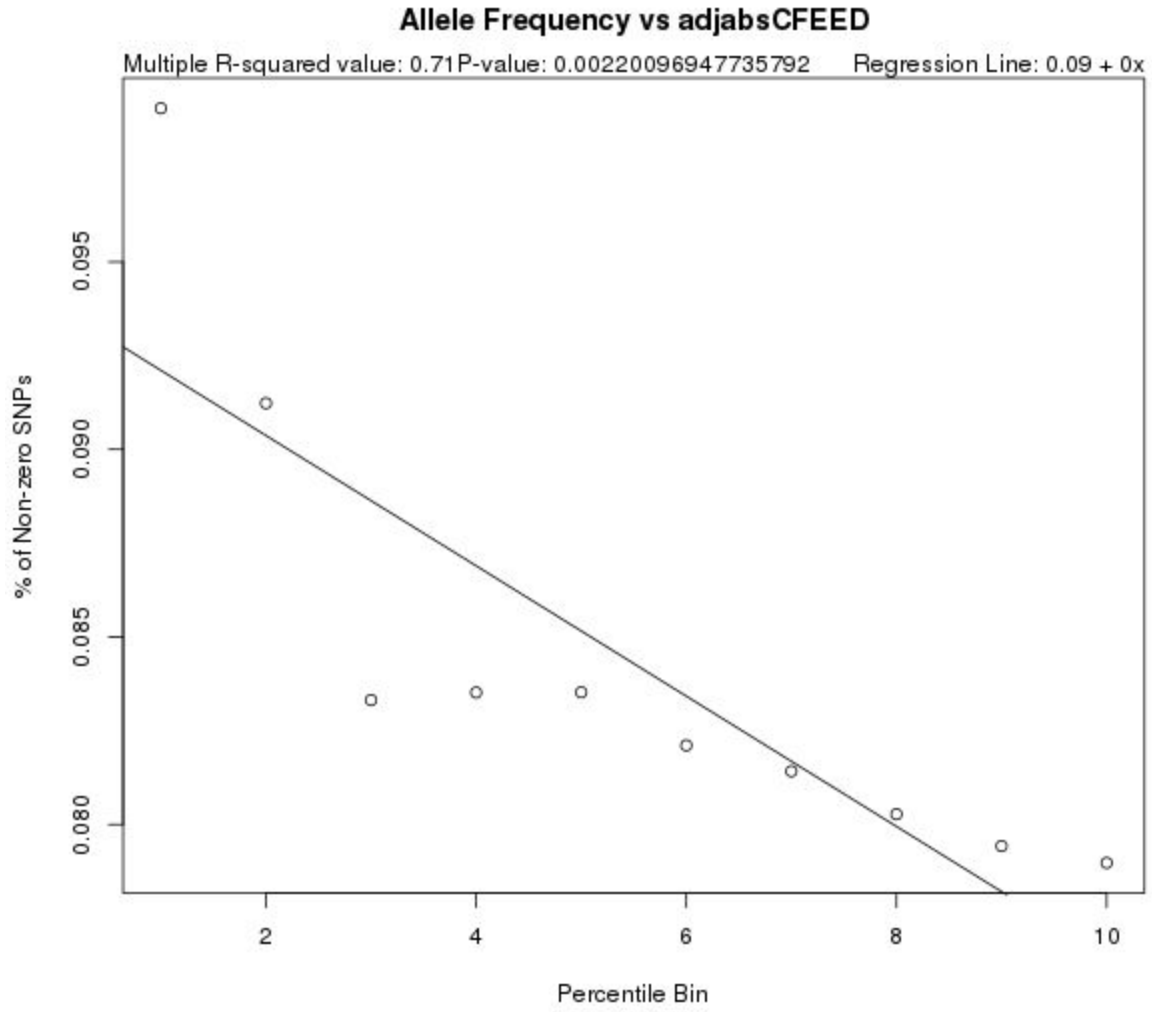
Supplementary Figure 97: Mean/Median GERP Score vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for Missense Variants



Supplementary Figure 98: % Non-zero Allele Frequency vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for Missense Variants



Supplementary Figure 99: Mean/Median GERP Score vs. Edit Distance Between Centroid Structures (CFEED) for Missense Variants



Supplementary Figure 100: % Non-zero Allele Frequency vs. Edit Distance Between Centroid Structures (CFEED) for Missense Variants

9. Figure and Table Appendix

- Figure 1: Diagram of the Overall Pipeline
- Figure 2: Example binned decile plots correlating EFEED disruptions to median GERP++ score (left) and population allele frequency (right)
- Figure 3: Percentage of disruption metrics for each RNA folding property that are significant for each constraint score for all SNPs
- Table 1: Percentage of disruption metrics for each RNA folding property that are significant for each constraint score for all SNPs; bolded cells indicate percentages greater than or equal to 50% and blue highlighted cells indicate percentages greater than or equal to 75%
- Figure 4: Percentage of disruption metrics for each RNA folding property that are significant for each constraint score for 5' UTR SNPs
- Table 2: Percentage of disruption metrics for each RNA folding property that are significant for each constraint score for 5' UTR SNPs; bolded cells indicate percentages greater than or equal to 50% and blue highlighted cells indicate percentages greater than or equal to 75%
- Figure 5: Percentage of disruption metrics for each RNA folding property that are significant for each constraint score for 3' UTR SNPs
- Table 3: Percentage of disruption metrics for each RNA folding property that are significant for each constraint score for 3' UTR SNPs; bolded cells indicate percentages greater than or equal to 50% and blue highlighted cells indicate percentages greater than or equal to 75%
- Figure 6: Percentage of disruption metrics for each RNA folding property that are significant for each constraint score for synonymous SNPs
- Table 4: Percentage of disruption metrics for each RNA folding property that are significant for each constraint score for synonymous SNPs; bolded cells indicate percentages greater than or equal to 50% and blue highlighted cells indicate percentages greater than or equal to 75%
- Figure 7: Percentage of disruption metrics for each RNA folding property that are significant for each constraint score for missense SNPs

- Table 5: Percentage of disruption metrics for each RNA folding property that are significant for each constraint score for missense SNPs; bolded cells indicate percentages greater than or equal to 50% and blue highlighted cells indicate percentages greater than or equal to 75%
- Figure 8: The user interface of "SNP mRNA Folding Consequences in Humans"
- Supplementary Figure 1: GERP Score vs. Change in Minimum Free Energy (dMFE) for All Transcript Regions
- Supplementary Figure 2: % Non-zero Allele Frequency vs. Change in Minimum Free Energy (dMFE) for All Transcript Regions
- Supplementary Figure 3: GERP Score vs. Change in Ensemble Free Energy (dEFE) for All Transcript Regions
- Supplementary Figure 4: % Non-zero Allele Frequency vs. Change in Ensemble Free Energy (dEFE) for All Transcript Regions
- Supplementary Figure 5: GERP Score vs. Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for All Transcript Regions
- Supplementary Figure 6: % Non-zero Allele Frequency vs. Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for All Transcript Regions
- Supplementary Figure 7: GERP Score vs. Change in Free Energy of the Centroid (dCFE) for All Transcript Regions
- Supplementary Figure 8: % Non-zero Allele Frequency vs. Change in Free Energy of the Centroid (dCFE) for All Transcript Regions
- Supplementary Figure 9: GERP Score vs. Change in Ensemble Diversity (dEND) for All Transcript Regions
- Supplementary Figure 10: % Non-zero Allele Frequency vs. Change in Ensemble Diversity (dEND) for All Transcript Regions
- Supplementary Figure 11: GERP Score vs. Change in Distance of the Ensemble of Structures to the Centroid (dCD) for All Transcript Regions
- Supplementary Figure 12: % Non-zero Allele Frequency vs. Change in Distance of the Ensemble of Structures to the Centroid (dCD) for All Transcript Regions
- Supplementary Figure 13: GERP Score vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for All Transcript Regions
- Supplementary Figure 14: % Non-zero Allele Frequency vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for All Transcript Regions
- Supplementary Figure 15: GERP Score vs. Edit Distance Between Ensembles (EFEED) for All Transcript Regions

- Supplementary Figure 16: % Non-zero Allele Frequency vs. Edit Distance Between Ensembles (EFEED) for All Transcript Regions
- Supplementary Figure 17: GERP Score vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for All Transcript Regions
- Supplementary Figure 18: % Non-zero Allele Frequency vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for All Transcript Regions
- Supplementary Figure 19: GERP Score vs. Edit Distance Between Centroid Structures (CFEED) for All Transcript Regions
- Supplementary Figure 20: % Non-zero Allele Frequency vs. Edit Distance Between Centroid Structures (CFEED) for All Transcript Regions
- Supplementary Figure 21: GERP Score vs. Change in Minimum Free Energy (dMFE) for 5' UTR Variants
- Supplementary Figure 22: % Non-zero Allele Frequency vs. Change in Minimum Free Energy (dMFE) for 5' UTR Variants
- Supplementary Figure 23: GERP Score vs. Change in Ensemble Free Energy (dEFE) for 5' UTR Variants
- Supplementary Figure 24: % Non-zero Allele Frequency vs. Change in Ensemble Free Energy (dEFE) for 5' UTR Variants
- Supplementary Figure 25: GERP Score vs. Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for 5' UTR Variants
- Supplementary Figure 26: % Non-zero Allele Frequency vs. Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for 5' UTR Variants
- Supplementary Figure 27: GERP Score vs. Change in Free Energy of the Centroid (dCFE) for 5' UTR Variants
- Supplementary Figure 28: % Non-zero Allele Frequency vs. Change in Free Energy of the Centroid (dCFE) for 5' UTR Variants
- Supplementary Figure 29: GERP Score vs. Change in Ensemble Diversity (dEND) for 5' UTR Variants
- Supplementary Figure 30: % Non-zero Allele Frequency vs. Change in Ensemble Diversity (dEND) for 5' UTR Variants
- Supplementary Figure 31: GERP Score vs. Change in Distance of the Ensemble of Structures to the Centroid (dCD) for 5' UTR Variants
- Supplementary Figure 32: % Non-zero Allele Frequency vs. Change in Distance of the Ensemble of Structures to the Centroid (dCD) for 5' UTR Variants
- Supplementary Figure 33: GERP Score vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for 5' UTR Variants

- Supplementary Figure 34: % Non-zero Allele Frequency vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for 5' UTR Variants
- Supplementary Figure 35: GERP Score vs. Edit Distance Between Ensembles (EFEED) for 5' UTR Variants
- Supplementary Figure 36: % Non-zero Allele Frequency vs. Edit Distance Between Ensembles (EFEED) for 5' UTR Variants
- Supplementary Figure 37: GERP Score vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for 5' UTR Variants
- Supplementary Figure 38: % Non-zero Allele Frequency vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for 5' UTR Variants
- Supplementary Figure 39: GERP Score vs. Edit Distance Between Centroid Structures (CFEED) for 5' UTR Variants
- Supplementary Figure 40: % Non-zero Allele Frequency vs. Edit Distance Between Centroid Structures (CFEED) for 5' UTR Variants
- Supplementary Figure 41: GERP Score vs. Change in Minimum Free Energy (dMFE) for 3' UTR Variants
- Supplementary Figure 42: % Non-zero Allele Frequency vs. Change in Minimum Free Energy (dMFE) for 3' UTR Variants
- Supplementary Figure 43: GERP Score vs. Change in Ensemble Free Energy (dEFE) for 3' UTR Variants
- Supplementary Figure 44: % Non-zero Allele Frequency vs. Change in Ensemble Free Energy (dEFE) for 3' UTR Variants
- Supplementary Figure 45: GERP Score vs. Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for 3' UTR Variants
- Supplementary Figure 46: % Non-zero Allele Frequency vs. Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for 3' UTR Variants
- Supplementary Figure 47: GERP Score vs. Change in Free Energy of the Centroid (dCFE) for 3' UTR Variants
- Supplementary Figure 48: % Non-zero Allele Frequency vs. Change in Free Energy of the Centroid (dCFE) for 3' UTR Variants
- Supplementary Figure 49: GERP Score vs. Change in Ensemble Diversity (dEND) for 3' UTR Variants
- Supplementary Figure 50: % Non-zero Allele Frequency vs. Change in Ensemble Diversity (dEND) for 3' UTR Variants
- Supplementary Figure 51: GERP Score vs. Change in Distance of the Ensemble of Structures to the Centroid (dCD) for 3' UTR Variants
- Supplementary Figure 52: % Non-zero Allele Frequency vs. Change in Distance of the Ensemble of Structures to the Centroid (dCD) for 3' UTR Variants

- Supplementary Figure 53: GERP Score vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for 3' UTR Variants
- Supplementary Figure 54: % Non-zero Allele Frequency vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for 3' UTR Variants
- Supplementary Figure 55: GERP Score vs. Edit Distance Between Ensembles (EFEED) for 3' UTR Variants
- Supplementary Figure 56: % Non-zero Allele Frequency vs. Edit Distance Between Ensembles (EFEED) for 3' UTR Variants
- Supplementary Figure 57: GERP Score vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for 3' UTR Variants
- Supplementary Figure 58: % Non-zero Allele Frequency vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for 3' UTR Variants
- Supplementary Figure 59: GERP Score vs. Edit Distance Between Centroid Structures (CFEED) for 3' UTR Variants
- Supplementary Figure 60: % Non-zero Allele Frequency vs. Edit Distance Between Centroid Structures (CFEED) for 3' UTR Variants
- Supplementary Figure 61: GERP Score vs. Change in Minimum Free Energy (dMFE) for Synonymous Variants
- Supplementary Figure 62: % Non-zero Allele Frequency vs. Change in Minimum Free Energy (dMFE) for Synonymous Variants
- Supplementary Figure 63: GERP Score vs. Change in Ensemble Free Energy (dEFE) for Synonymous Variants
- Supplementary Figure 64: % Non-zero Allele Frequency vs. Change in Ensemble Free Energy (dEFE) for Synonymous Variants
- Supplementary Figure 65: GERP Score vs. Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for Synonymous Variants
- Supplementary Figure 66: % Non-zero Allele Frequency vs. Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for Synonymous Variants
- Supplementary Figure 67: GERP Score vs. Change in Free Energy of the Centroid (dCFE) for Synonymous Variants
- Supplementary Figure 68: % Non-zero Allele Frequency vs. Change in Free Energy of the Centroid (dCFE) for Synonymous Variants
- Supplementary Figure 69: GERP Score vs. Change in Ensemble Diversity (dEND) for Synonymous Variants
- Supplementary Figure 70: % Non-zero Allele Frequency vs. Change in Ensemble Diversity (dEND) for Synonymous Variants

- Supplementary Figure 71: GERP Score vs. Change in Distance of the Ensemble of Structures to the Centroid (dCD) for Synonymous Variants
- Supplementary Figure 72: % Non-zero Allele Frequency vs. Change in Distance of the Ensemble of Structures to the Centroid (dCD) for Synonymous Variants
- Supplementary Figure 73: GERP Score vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for Synonymous Variants
- Supplementary Figure 74: % Non-zero Allele Frequency vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for Synonymous Variants
- Supplementary Figure 75: GERP Score vs. Edit Distance Between Ensembles (EFEED) for Synonymous Variants
- Supplementary Figure 76: % Non-zero Allele Frequency vs. Edit Distance Between Ensembles (EFEED) for Synonymous Variants
- Supplementary Figure 77: GERP Score vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for Synonymous Variants
- Supplementary Figure 78: % Non-zero Allele Frequency vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for Synonymous Variants
- Supplementary Figure 79: GERP Score vs. Edit Distance Between Centroid Structures (CFEED) for Synonymous Variants
- Supplementary Figure 80: % Non-zero Allele Frequency vs. Edit Distance Between Centroid Structures (CFEED) for Synonymous Variants
- Supplementary Figure 81: GERP Score vs. Change in Minimum Free Energy (dMFE) for Missense Variants
- Supplementary Figure 82: % Non-zero Allele Frequency vs. Change in Minimum Free Energy (dMFE) for Missense Variants
- Supplementary Figure 83: GERP Score vs. Change in Ensemble Free Energy (dEFE) for Missense Variants
- Supplementary Figure 84: % Non-zero Allele Frequency vs. Change in Ensemble Free Energy (dEFE) for Missense Variants
- Supplementary Figure 85: GERP Score vs. Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for Missense Variants
- Supplementary Figure 86: % Non-zero Allele Frequency vs. Change in Free Energy of the Maximum Expected Accuracy Structure (dMEAFE) for Missense Variants
- Supplementary Figure 87: GERP Score vs. Change in Free Energy of the Centroid (dCFE) for Missense Variants

- Supplementary Figure 88: % Non-zero Allele Frequency vs. Change in Free Energy of the Centroid (dCFE) for Missense Variants
- Supplementary Figure 89: GERP Score vs. Change in Ensemble Diversity (dEND) for Missense Variants
- Supplementary Figure 90: % Non-zero Allele Frequency vs. Change in Ensemble Diversity (dEND) for Missense Variants
- Supplementary Figure 91: GERP Score vs. Change in Distance of the Ensemble of Structures to the Centroid (dCD) for Missense Variants
- Supplementary Figure 92: % Non-zero Allele Frequency vs. Change in Distance of the Ensemble of Structures to the Centroid (dCD) for Missense Variants
- Supplementary Figure 93: GERP Score vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for Missense Variants
- Supplementary Figure 94: % Non-zero Allele Frequency vs. Edit Distance Between Minimum Free Energy Structures (MFEED) for Missense Variants
- Supplementary Figure 95: GERP Score vs. Edit Distance Between Ensembles (EFEED) for Missense Variants
- Supplementary Figure 96: % Non-zero Allele Frequency vs. Edit Distance Between Ensembles (EFEED) for Missense Variants
- Supplementary Figure 97: GERP Score vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for Missense Variants
- Supplementary Figure 98: % Non-zero Allele Frequency vs. Edit Distance Between Maximum Expected Accuracy Structures (MEAED) for Missense Variants
- Supplementary Figure 99: GERP Score vs. Edit Distance Between Centroid Structures (CFEED) for Missense Variants
- Supplementary Figure 100: % Non-zero Allele Frequency vs. Edit Distance Between Centroid Structures (CFEED) for Missense Variants